genes

MDPI

Article

# Identification of Regulatory SNPs Associated with Vicine and Convicine Content of *Vicia faba* Based on Genotyping by Sequencing Data Using Deep Learning

**Felix Heinrich [1], Martin Wutke [1], Pronaya Prosun Das [1], Miriam Kamp [1], Mehmet Gültas [1,2], Wolfgang Link [3] and Armin Otto Schmitt [1,2,*]**

1 Breeding Informatics Group, Department of Animal Sciences, Georg-August University, Margarethe von Wrangell-Weg 7, 37075 Göttingen, Germany; felix.heinrich@uni-goettingen.de (F.H.); martin.wutke@uni-goettingen.de (M.W.); pronayaprosun.das@stud.uni-goettingen.de (P.P.D.); miriam-kamp@gmx.net (M.K.); gueltas@cs.uni-goettingen.de (M.G.)

2 Center for Integrated Breeding Research (CiBreed), Albrecht-Thaer-Weg 3, Georg-August University, 37075 Göttingen, Germany

3 Department of Crop Sciences, Georg-August University, Von-Siebold-Str. 8, 37075 Göttingen, Germany; wlink@gwdg.de

* Correspondence: armin.schmitt@uni-goettingen.de

Received: 29 April 2020; Accepted: 28 May 2020; Published: 5 June 2020


**Abstract:** Faba bean (*Vicia faba*) is a grain legume, which is globally grown for both human consumption as well as feed for livestock. Despite its agro-ecological importance the usage of *Vicia faba* is severely hampered by its anti-nutritive seed-compounds vicine and convicine (V+C). The genes responsible for a low V+C content have not yet been identified. In this study, we aim to computationally identify regulatory SNPs (rSNPs), i.e., SNPs in promoter regions of genes that are deemed to govern the V+C content of *Vicia faba*. For this purpose we first trained a deep learning model with the gene annotations of seven related species of the Leguminosae family. Applying our model, we predicted putative promoters in a partial genome of *Vicia faba* that we assembled from genotyping-by-sequencing (GBS) data. Exploiting the synteny between *Medicago truncatula* and *Vicia faba*, we identified two rSNPs which are statistically significantly associated with V+C content. In particular, the allele substitutions regarding these rSNPs result in dramatic changes of the binding sites of the transcription factors (TFs) MYB4, MYB61, and SQUA. The knowledge about TFs and their rSNPs may enhance our understanding of the regulatory programs controlling V+C content of *Vicia faba* and could provide new hypotheses for future breeding programs.

**Keywords:** promoter; rSNP; convolutional neural network; *Vicia faba*; GBS; vicin/convicin

## 1. Introduction

New methods in the field of genome sequencing—commonly summarized as next generation sequencing (NGS)—offer cost-effective strategies to produce massive amounts of sequencing data. One of these methods is genotyping-by-sequencing (GBS), which is an efficient method to obtain genome-wide genotype data for any species [1]. The characteristic feature of GBS is the reproducible generation of short genomic fragments using known restriction enzymes. Thanks to its easy applicability, GBS is currently the method of choice in the field of plant sciences since it makes plants without reference genome amenable to genomic analysis. Several groups have applied GBS to obtain high-quality genome-wide SNP markers. These markers have often been used for applications

like GWAS, marker-assisted selection, breeding value estimation in genomic prediction, analysis of high density genetic maps, or assessment of population dynamics in plant genomics and plant breeding [2–10]. Another remarkable feature of GBS has been highlighted by Elshire et al. [11], namely that it made the analysis of regulatory regions of genes such as promoters feasible. Despite the growing interest in the analysis of GBS sequenced reads, this capacity of GBS has been poorly studied. This sequencing approach is particularly important for crop species which still often lack a reference genome sequence such as *Vicia faba*. The capacity of GBS to generate sequences of regulatory regions could be used for the prediction of such regions, e.g., promoters, in *Vicia faba*.

The faba bean is an Old World grain legume, which is grown both for combine harvested feed and as vegetable crop for human consumption. It is diploid with $2x = 12$ very large chromosomes. Due to its large size of 13 Gbp [12], there is thus far no sequenced and annotated reference genome available for this plant. Despite its agro-ecological importance (N-symbiosis, rotation hygiene, and pollinator support) [13] it is a crop of limited importance in many countries. This is mainly caused by its anti-nutritive seed-compounds vicine and convicine, which are co-occurring pyrimidine glycosides (in the following termed V+C) and have negative effects to animals such as laying hens, broilers and piglets, but also to 400 million humans suffering from G6PD deficiency [14,15]. The V+C content is a factor that severely limits the wider usage of *Vicia faba* as feed for animals and food for humans. Breeding V+C-poor varieties and production and marketing of their fruits could have a range of positive effects including e.g., reduction of environmentally critical soya bean imports into Europe and Northern America, fostering of regional production methods, and avoidance of energy intensive transports. To date, the location of the gene controlling the V+C content could only be restricted to a region on chromosome 1 of *Vicia faba* that exhibits conserved synteny with chromosome 2 of the related species *Medicago truncatula* between the *Medicago truncatula* genes Medtr2g008210 and Medtr2g010180 [14,16].

The promoter of a gene is the region immediately around its transcription start site (TSS) as well as further upstream of it. The promoter contains multiple elements that allow the binding of the RNA polymerase II (Pol II) along with transcription factors (TFs), thus controlling the transcription of the associated gene. Due to their impact on the gene regulation the SNPs located in the promoter regions that affect the transcription factor binding sites (TFBSs) are commonly called regulatory SNPs (rSNPs). Today it is well known that these rSNPs may be causal for the phenotype and could therefore possibly provide prime candidates useful for breeding programs or marker-assisted selection [17–22]. Despite the rich literature on the analysis of promoters, their prediction remains a challenging task due to their complex and diverse structure. Until now, different machine learning approaches have been developed, which form the core of most computational prediction methods for promoter regions. Whereas in early works the emphasis was on the identification of specific promoter elements (such as TATA boxes, initiator elements (Inrs), downstream promoter elements (DPE) and others) or extraction of k-mer distributions [23–30], nowadays a more holistic approach is given preference in that whole genomic regions are examined in Convolutional Neural Networks (CNNs), which have been successfully applied in many species [31–36].

A large scale genome-wide key study has been conducted by Kumari et al. for the prediction and analysis of core promoter elements (CPEs) across plant monocots and dicots [37]. For this purpose, CPEs of four monocots and four dicots were comprehensively analyzed and compared to establish the common as well as the specific properties of CPEs in promoter sequences. The results obtained in [37] are, on the one hand, promising to enhance the limited knowledge available about the differences between dicots and monocots with respect to their CPEs. On the other hand, they contributed to gain novel insight into the plant promoter sequence architectures and showed that some promoter signatures are strongly conserved within larger groups of plants like monocots or dicots. Based on these findings, Shahmuradov et al. developed a model, namely TSSPlant, for the prediction of plant Pol II promoters across species boundaries [38].

In line with the studies of Kumari et al. [37] and Shahmuradov et al. [38] we designed an analysis workflow for the prediction of promoter sequences as well as rSNPs of *Vicia faba* in this study. For this purpose, we first trained a CNN model using the known promoter sequences of seven plants of the Leguminosae family. Second, using GBS sequence reads of 20 *Vicia faba* lines with known V+C content, we assembled a *de novo* draft partial genome. Thereafter, we called the genomic variants by aligning the GBS reads to the partial genome to obtain high quality SNPs for candidate gene association studies. Next, applying our CNN model to the partial genome sequences, we have predicted the potential promoter sequences of *Vicia faba*. Finally, we analyzed the SNPs in these promoter sequences that were associated with the V+C content of *Vicia faba* regarding their effect on the binding affinity of TFs. Our results show that 2.46% of the assembled sequences were predicted to be promoters. We found 14 regulatory SNPs that could be mapped to the syntenic *Medicago truncatula* region harbouring the major gene for low V+C content [14,16]. These findings could be of use to increase our understanding of the regulation of the V+C content and could provide novel genomic targets for future breeding strategies of V+C poor *Vicia faba* varieties.

## 2. Materials and Methods

### 2.1. Plant Material and Sequencing

In total, 20 inbred lines of *Vicia faba* were selected of which six had low V+C content and 14 had high V+C content (see Supplementary Table S5). The lines were inbred via single-seed descent from cultivars, from a gene-bank accession, from biparental crosses or from a landrace and include winter and spring types. DNA was extracted from the grains of the plants. Two pooled grains were used per line. DNA extraction was done with LGC's sbeadex livestock kit following the lysis protocol L for plant tissue. Genotyping-by-sequencing was carried out on the Illumina NextSeq 500 V2 platform. The DNA was digested with the restriction enzyme MslI (recognition sequence: CAYNN^NNRTG). Per sample ~3 million 150 bp paired end reads were obtained. Then sequencing adapter remnants were clipped and reads whose 5′ ends did not match the restriction enzyme site were discarded. The sequencing and filtering was performed by LGC Genomics GmbH (Berlin, Germany).

### 2.2. Assembly of a Partial Vicia faba Genome

Following the *de novo assembly* strategies used in [39,40], we applied the *de novo* assembler Trinity [41] to the GBS sequence reads for the construction of a partial genome for *Vicia faba*. In total, 694,605 contigs with an average length of 236 bp were constructed. To filter out redundant contigs, we clustered the contigs with `CD-HIT` [42] using a threshold of 95.0% for sequence identity.

### 2.3. Variant Calling and Association Testing

Following the variant calling pipeline outlined in [43], we mapped the sequence reads onto the partial genome using `Bowtie2` [44]. The variant calling was done with `SAMtools mpileup` [45]. We excluded structural variants such as insertions and deletions as well as non-biallelic SNPs yielding 1,880,592 SNPs. Low quality SNPs with a quality score of lower than 400 were excluded using `PLINK 1.9` [46]. The association between candidate SNPs and the V+C content was tested with `PLINK` using a 1df chi-squared allelic test. To control the type I error rate we set the false discovery rate (FDR) to 0.1.

### 2.4. Data Sets for Training the Neural Network

Mainly considering the members of the Leguminosae family, we used in our analysis seven species (*Glycine max*, *Lupinus angustifolius*, *Medicago truncatula*, *Phaseolus vulgaris*, *Trifolium pratense*, *Vigna angularis*, and *Vigna radiata*) that have a complete and annotated reference genome sequence available. To further establish the cross-species promoter prediction performance of our CNN model with a more distant plant, we also chose to include in our analysis the model species *Arabidopsis thaliana*. Following [31,33], we extracted for each species their core promoter sequences covering

the −200 bp to +50 bp regions relative to the transcription start sites (TSSs) of protein coding genes from the Ensembl Plants database (release 45) [47] using BioMart [48]. Simultaneously, the sequences covering [TSS+751,TSS+1000] from the core gene region of the genes were extracted, as non-promoter sequences. Sequences that were not assigned to a chromosome or which contained ambiguous bases were not considered.

Currently, due to the absence of an annotated reference genome, there is only scarce knowledge about the promoter sequence architecture in *Vicia faba*. Hence, it is still challenging to determine characteristic signatures of *Vicia faba* promoters that distinguish them from non-promoter regions. To eliminate this lack of knowledge to some extent and to enhance the distinction of promoters vs. non-promoters in *Vicia faba*, the consideration of additional non-promoter sets is important. Consequently, we included two further sets of sequences of length 250 bp as non-promoters in our analysis. While the first set was randomly extracted from the *Medicago truncatula* reference genome by excluding the region [TSS-1000,TSS+500], the second set was sampled from the *Vicia faba* reference transcriptome V2 which was downloaded from the Pulse Crop Database [49]. The final number of sequences for each data set can be found in Table 1.

**Table 1.** Number of promoter and non-promoter sequences in the sets that were used as training sets.

| Species | # Promoter Sequences | # Non-Promoter Sequences |
|---|---|---|
| *Arabidopsis thaliana* | 23,315 | 23,315 |
| *Glycine max* | 46,199 | 46,199 |
| *Lupinus angustifolius* | 23,463 | 23,463 |
| *Medicago truncatula* | 32,158 | 32,158 |
| *Phaseolus vulgaris* | 22,750 | 22,750 |
| *Trifolium pratense* | 14,749 | 14,749 |
| *Vigna angularis* | 19,584 | 19,584 |
| *Vigna radiata* | 15,495 | 15,495 |
| *Medicago truncatula* (Genome-wide) | - | ∼12,000 |
| *Vicia faba* (Transcriptome) | - | ∼60,000 |

### 2.5. Sequence Features Used to Predict Promoters

In line with previous studies [33,38], we used a variety of additional features to characterize promoter sequences as precisely as possible. We have determined the distribution of the following features for the sequence sets listed in Table 1:

**Feature 1:** Frequency of the dinucleotides CA and CG

**Feature 2:** Frequency of the TATA motif

**Feature 3:** CG-skew of sequences ($CG_{\text{skew}} = \frac{\#C-\#G}{\#C+\#G}$ where $\#C$ and $\#G$ refer to the counts of nucleotides $C$ and $G$ in the sequences)

**Feature 4:** Frequency of $k$-mers using different values for $k$

**Information theory based features:** we included in our analysis two additional features, namely the Horizontal Mutual Information (HMI) and the Generalized Topological Entropy (GTE).

The HMI is calculated based on a predefined distance $d$ between two positions in a sequence and provides a measure of auto-covariation between the nucleotides of interest [50].

$$\text{HMI}(d) = \sum_{m=\{A,C,G,T\}} \sum_{n=\{A,C,G,T\}} p_{mn}(d) \cdot \log \frac{p_{mn}(d)}{p_m(d)p_n(d)}, \quad (1)$$

where $p_m(d)$, $p_n(d)$ and $p_{mn}(d)$ refer to the marginal and joint probabilities of the nucleotides being $d$ bp apart. A high value of HMI($d$) indicates a strong correlation between the nucleotides regarding their distance $d$.

Entropy is a measure to reflect the complexity of sequences. It has been used to characterize the randomness of DNA sequences [51]. More specific varieties of entropies such as the GTE have been successfully applied in [52,53] to explore and to compare the complexity of introns, exons,

and promoter regions. Based on the findings of these studies, we included GTE as an additional feature in our analysis which could provide an important information.

Let $\omega$ be a DNA sequence of length $|\omega|$ and let $n_\omega$ be the unique integer such that $4^n + n - 1 \leq |\omega| < 4^{n+1} + (n+1) + 1$. Then the GTE is defined as

$$H^k_{n_\omega}(\omega) = \frac{1}{k} \sum_{i=n_\omega-k+1}^{n_\omega} \frac{\log_4(p_\omega(i))}{i} \tag{2}$$

where $p_\omega(i)$ refers to the number of unique sub-sequences of length $i$ that appear in $\omega$. We set $k = n_\omega$ to consider sub-sequences of all possible lengths.

### *2.6. Convolutional Neural Networks*

Our proposed model follows a CNN architecture, which is nowadays one of the most popular neural network architectures [54]. Using convolutional layers as its core elements, a CNN is able to automatically learn local as well as global features from the data layer-wise by applying a convolution operation and by encoding specific aspects of the data [55–57]. Within a layer, an array of stacked weight matrices of dimension $W \times H \times D$, where $W$, $H$, and $D$ correspond to the width, height, and depth of the array, respectively, is moved spatially across the input data [31,34]. At every possible position, the summed element-wise product between the weight matrices and a subset of the input is calculated and a corresponding feature map is computed.

The structure of the network that was used is illustrated in Figure 1. The input of the network is formed by a sequence of nucleotides of length 250 bp where each nucleotide is encoded into a one-hot representation and expressed by a four-dimensional vector, with A encoded as $(1,0,0,0)$, C as $(0,1,0,0)$, G as $(0,0,1,0)$, and T as $(0,0,0,1)$. As can be seen in Figure 1, the network is composed of four 1D-convolutional layers followed by a flattening layer, two fully-connected layers and an output layer. All convolutional layers are implemented using a ReLU activation, a stride parameter of 2, zero-padding and a filter size of 21. The first layer uses 64 filters, whereas the second, third and fourth layers use 128, 256, and 512 filters, respectively. To avoid overfitting the training data, a dropout layer with rate $= 0.2$ is used after each convolution [58]. After processing of the sequences by the convolutional layers, a flattening layer transforms the output to a one-dimensional vector and passes its values to two consecutive fully-connected layers with 128 and 64 neurons, respectively. Finally, an output layer with a sigmoid activation classifies the input sequence as promoter or non-promoter. Additional features were included one-by-one by concatenating their values to the flattening layer in order to explore their effect on the improvement of the classification performance.

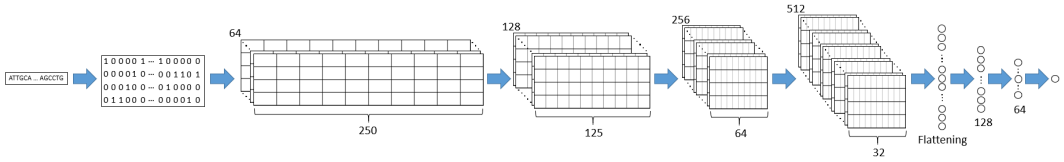


**Figure 1.** The network architecture of the CNN promoter prediction consists of four 1D-convolutional layers followed by a flattening layer and two fully-connected layers. At the end, an output layer with one neuron and a sigmoid activation function computes the probability that the analyzed sequence is classified as a promoter sequence.

Before training the final model, a separate network for each species was trained individually using the Adam optimizer [59], L2-regularization and binary cross-entropy loss [60]. For each network, 90% of the sequences were used for model training and 10% for testing. The CNN was implemented in R using Keras [61] with TensorFlow [62] as a backend.

To assess the prediction performance we identified the number of correctly predicted promoter and non-promoter sequences as True Positives (TP) and True Negatives (TN), as well as the number of

true promoter sequences predicted as non-promoter sequences, False Negatives (FN), and the number of true non-promoter sequences predicted as promoter sequences, False Positives (FP). From these measures, we calculated Accuracy (ACC), Sensitivity (true positive rate), Specificity (true negative rate), and the Matthews Correlation Coefficient (MCC) as below [33,34,63]:

$$\text{ACC} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{5}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

### 2.7. Identification of Putative Regulatory SNPs

In order to identify regulatory SNPs (rSNPs), we analyzed the predicted promoter sequences of *Vicia faba*. For this purpose, we first selected all SNPs that are located in promoters and that we could successfully map against the *Medicago truncatula* genome from the initial 685,215 SNPs (see Section 2.3). Second, we extracted for each SNP its flanking sequence covering ±25 bp relative to the position of the SNP. Third, two copies of the extracted sequences were created: while the first sequence contained at the SNP position the reference allele, the second contained the alternate allele. Thereafter, we identified putative transcription factor binding sites (TFBSs) by applying the MATCH$^{TM}$ program [64] together with a non-redundant plant position weight matrix (PWM) library obtained from the TRANSFAC database [65] to the flanking sequences of each SNP. The MATCH$^{TM}$ program provides for each putative TFBS a matrix similarity score (MSS) ranging from zero to one, which reflects the potential binding affinity of the related TF to it. Finally, we predicted the consequence of each SNP on the TFBS by comparing their MSSs in the two sequences. As a result we observed in our analysis four different types of consequences: (i) no effect, (ii) change in binding affinity, (iii) loss of TFBS (a TFBS appears only for the reference allele) and (iv) gain of TFBS (a TFBS appears only for the alternate allele). Two TFBSs are considered as identical if their PWMs, positions and their strands are equal for both alleles. If the scores computed by MATCH$^{TM}$ are identical in both alleles, the SNP is assumed to have no effect on the TFBS. In our further analysis, we define a SNP as a rSNP, if it has an effect on the binding affinity of at least one TF, i.e., if its type of consequence is (ii), (iii), or (iv).

## 3. Results and Discussion

Classical application of GBS includes the identification and genotyping of large numbers of genomic variants. This provides several possibilities in plant breeding like the discovery of important markers by GWAS even in the absence of the reference genome. In this study, however, we focused on another important property of the GBS approach, namely its capacity to access regulatory regions (especially promoters) which serves as a basis for the identification of rSNPs in *Vicia faba*.

### 3.1. Processing the GBS Data

Sequencing of the 20 *Vicia faba* samples yielded 51 GB of GBS data. The *de novo assembly* and filtering resulted in a partial genome consisting of 419,390 contigs with a total length of 100,037,292 bp. Considering that the proposed size of the *Vicia faba* genome is about 13 Gbp [12] our partial genome covered 0.77% of the total genome. Through remapping of the reads to the partial genome with `Bowtie2` and subsequent variant calling with `SAMtools` 1,880,592 SNPs could be derived. The quality scores of these SNPs as given in the vcf file showed a clear bimodal distribution with a minimum

at a quality score of 400. 1,195,377 SNPs having a quality score of lower than 400 were discarded, such that 685,215 high quality SNPs remained.

### 3.2. Prediction of Promoter Sequences

#### 3.2.1. Intra- and Inter-Species Promoter Prediction

In order to gain first insights into the predictability of promoters of the seven Leguminosae family members and *Arabidopsis thaliana*, we trained our CNN model for each species individually. The prediction reliability of the CNN model has been examined for each species by classifying the intra- and inter-species promoters that were not used in the training process. To assess the performance of the classification, the ACC, Sensitivity, Specificity and MCC values were calculated. The details based on the ACC values are presented in Table 2 and the results based on the remaining measures are given in the Supplementary Tables S1–S3.

**Table 2.** ACC values of the intra- and inter-species promoter classification using the species-specific trained CNNs. Off-diagonal numbers are ACC values for inter-species classification, diagonal numbers are ACC values for intra-species classification. For instance, a CNN trained on *Lupinus angustifolius* and used for classification of *Vigna angularis* promoters has an accuracy of 0.974.

| Evaluated / Trained | *Arabidopsis thaliana* | *Glycine max* | *Lupinus angustifolius* | *Medicago truncatula* | *Phaseolus vulgaris* | *Trifolium pratense* | *Vigna angularis* | *Vigna radiata* |
|---|---|---|---|---|---|---|---|---|
| *Arabidopsis thaliana* | 0.901 | 0.767 | 0.690 | 0.746 | 0.797 | 0.765 | 0.633 | 0.733 |
| *Glycine max* | 0.837 | 0.864 | 0.915 | 0.847 | 0.863 | 0.724 | 0.914 | 0.856 |
| *Lupinus angustifolius* | 0.545 | 0.611 | 0.981 | 0.720 | 0.586 | 0.493 | 0.974 | 0.709 |
| *Medicago truncatula* | 0.755 | 0.797 | 0.959 | 0.876 | 0.789 | 0.715 | 0.951 | 0.841 |
| *Phaseolus vulgaris* | 0.845 | 0.842 | 0.888 | 0.834 | 0.898 | 0.748 | 0.880 | 0.853 |
| *Trifolium pratense* | 0.822 | 0.764 | 0.696 | 0.751 | 0.794 | 0.840 | 0.689 | 0.736 |
| *Vigna angularis* | 0.510 | 0.607 | 0.971 | 0.715 | 0.583 | 0.494 | 0.977 | 0.712 |
| *Vigna radiata* | 0.741 | 0.812 | 0.937 | 0.827 | 0.825 | 0.675 | 0.928 | 0.904 |

The results presented in Table 2 show that although the CNN models have been trained only using one-hot representation of sequences for each species individually, the network architecture is able to recognize certain patterns in the sequences which leads to the predictability of promoters across different species to a certain degree. These findings support the results presented in [37] and indicate that some of the promoter signatures seem to be conserved between the *Leguminosae* family members.

Further, Table 2 demonstrates that the classification performance of some CNN models remarkably results in higher ACC values for inter-species prediction than for intra-species prediction. In particular, this is the case for the species *Lupinus angustifolius* and *Vigna angularis* whose promoters have been predicted with very high accuracy by the CNN models of the other species except for the *Arabidopsis thaliana* and *Trifolium pratense* models. This could be attributed to the underlying genome annotations of these species since their annotations seem to be created based on the genome information of well-studied family members. Especially, this assumption is true regarding the inclusion of the Leguminosae family specific promoter patterns in the promoters of these species. This hypothesis has been supported by the prediction performance of the *Lupinus angustifolius* model on the other species, which reached very high degrees of specificity while only achieving low degrees of sensitivity. To this end, we compared the inter-species prediction ACC values of the species regarding their performance on *Lupinus angustifolius* and *Vigna angularis*. This comparison revealed that the promoters of *Lupinus*

*angustifolius* were predicted with a slightly higher mean accuracy value than the promoters of *Vigna angularis* (0.880 compared to 0.868).

So far, we trained our CNN model only using the order of the nucleotides in DNA sequences (one-hot encoding). However, previous studies pointed out that the combination of one-hot encoding with additional widely used features could lead to a substantially improved performance in promoter identification [33,34]. For this purpose, we followed a similar procedure as suggested by Triska et al. in [33] and systematically evaluated the combination of sequence features (see Section 2.5) in the CNN model training of each species. The results are presented only for *Medicago truncatula* as an example in Table 3.

**Table 3.** Contribution of additional features in the CNN model of *Medicago truncatula*.

| Features | Accuracy | Sensitivity | Specificity | MCC |
|---|---|---|---|---|
| DNA sequence | 0.876 | 0.897 | 0.855 | 0.750 |
| DNA sequence + 2-mer | 0.874 | 0.880 | 0.867 | 0.747 |
| DNA sequence + 2-mer + frequency of CA motif | 0.862 | 0.828 | 0.897 | 0.726 |
| DNA sequence + 2-mer + frequency of CG motif | 0.875 | 0.875 | 0.876 | 0.751 |
| DNA sequence + 2-mer + HMI | 0.874 | 0.882 | 0.865 | 0.747 |
| DNA sequence + 2-mer + frequency of TATAA motif | 0.876 | 0.875 | 0.878 | 0.752 |
| DNA sequence + 2-mer + CG skew | 0.876 | 0.889 | 0.863 | 0.753 |
| DNA sequence + topological entropy | 0.874 | 0.886 | 0.861 | 0.747 |
| DNA sequence + 2-mer + topological entropy | 0.871 | 0.852 | 0.890 | 0.743 |
| DNA sequence + 2-mer + HMI + frequency of TATAA motif | 0.871 | 0.869 | 0.874 | 0.743 |
| DNA sequence + 2-mer + HMI + frequency of CA motif + frequency of CG motif + frequency of TATAA motif + CC skew | 0.873 | 0.859 | 0.888 | 0.747 |
| DNA sequence + HMI + frequency of CA motif + frequency of CG motif + frequency of TATAA motif + CG skew | 0.875 | 0.889 | 0.860 | 0.749 |

In contrast to previous studies [33,34], Table 3 shows that regardless of the usage of any additional feature, the performance of the CNN model could in general not be significantly improved. However, these results are in agreement with findings presented in [31,32,35,36] and indicate that the CNN architecture is able to learn specific patterns inherent in the sequences automatically. Hence, these patterns carry information which is obviously redundant to these widely used features. Consequently, it turns out that the consideration of additional features does not lead to an improvement of the CNN model performance and may, on the contrary, increase the noise during training.

#### 3.2.2. Prediction of *Vicia faba* Promoters

The knowledge about the promoter signatures which are conserved between the *Leguminosae* family members provides an important clue for the precise prediction of *Vicia faba* promoters, which still remains a challenge. However, the consideration of the sequences of only one *Leguminosae* family member in the CNN model could be insufficient to capture the variety of different promoter signatures for the accurate computational identification of the *Vicia faba* promoters. To mitigate the drawback of single species models we systematically examined different CNN models seeking to determine the preferential combination of *Leguminosae* family members by intensifying the signal of promoter sequences and thus to improve the performance of the CNN model. Consequently, we trained a CNN model based on the species *Lupinus angustifolius* and *Medicago truncatula* since the combined usage of their manually selected sequences perfectly complement each other. In the last step, we included in the CNN model training two additional non-promoter sets (defined in Section 2.4) to enhance the

distinction signals between promoter and non-promoter regions. The training sequences are given in the Supplementary Files S6 and S7. The evaluation of this CNN model yields to clearly better ACC and MCC values of 0.98 and 0.95, respectively. A further analysis reveals that the usage of other sequence features together with one-hot encoding in our final CNN model does not affect the performance of the classifier.

Finally, by applying the CNN model to the *Vicia faba* sequences of length 250 bp, we classified in total 2.46% of them as potential promoter sequences. It is important to note that, due to the random fragment orientation regarding the direction of the reads from GBS, the correct direction of the sequences in the *de novo* assembly draft partial genome of *Vicia faba* is unknown. To address this limitation, we considered in our predictions four different types of the sequences as: (i) the original obtained assembly; (ii) the complement of the obtained assembly that is gained by keeping the reading direction; (iii) the reverse of the obtained assembly that is gained by changing the reading direction; and (iv) the reverse complement of the obtained assembly.

Checking the positions of the SNPs in the contigs disclosed that in total 132,399 out of 685,215 SNPs were located in the predicted *Vicia faba* promoters. A flanking sequence of $\pm 25$ bp could only be obtained for 118,492 SNPs. These SNPs with a complete flanking sequence were mapped against the *Medicago truncatula* genome using the BLASTN algorithm with a threshold of 0.01 for the *e-value* and of 0.9 for the *percent identity* [66]. Overall, we found 33,846 hits for 1976 SNPs showcasing the repetitiveness of the *Medicago truncatula* genome. We identified 14 SNPs that map to the predefined target region of *Medicago truncatula* that harbours orthologous genes associated with the V+C content of *Vicia faba* [14,16]. This target region is ranging approximately from 1,300,000 bp to 2,300,000 bp of the *Medicago truncatula* chromosome 2. An overview of these SNPs and their mapped position in the *Medicago truncatula* genome along with the genes with the closest TSS is given in Table 4. We tested these 14 SNPs for their association with the V+C content with `PLINK`. The adjusted *p*-values presented in Table 4 suggest that SNP_341016_236 and SNP_341016_239, which are located in the same promoter, show a highly significant association with the V+C content in *Vicia faba* while the associations of the remaining SNPs are not significant at the level $\alpha = 0.05$. For both of these SNPs the reference allele only occurs in the low V+C lines with one exception while the alternate allele is restricted to the high V+C lines (see Supplementary Table S8).

**Table 4.** The 14 SNPs found in the predicted promoters of *Vicia faba* that were mapped to the *Medicago truncatula* target genomic region.

| SNP_ID | Genotype | FDR | Position | Medicago Gene |
|---|---|---|---|---|
| SNP_131938_118 | C/T | 0.234 | 1,385,390 | MTR_2g008290 |
| SNP_302904_183 | G/A | 0.179 | 1,385,444 | |
| SNP_341016_236 | C/T | $1.17 \cdot 10^{-7}$ | 1,554,857 | MTR_2g008620 |
| SNP_341016_239 | G/A | $1.17 \cdot 10^{-7}$ | 1,554,860 | |
| SNP_356745_200 | A/G | 0.730 | 1,707,078 | MTR_2g008960 |
| SNP_280549_41 | C/T | 0.234 | 1,707,183 | |
| SNP_350273_103 | G/T | 0.234 | 1,707,199 | |
| SNP_350273_90 | A/C | 0.234 | 1,707,212 | |
| SNP_350273_61 | G/A | 0.234 | 1,912,704 | MTR_2g009430 |
| SNP_29452_204 | G/A | 0.730 | 1,912,812 | |
| SNP_29452_206 | G/A | 0.496 | 1,912,814 | |
| SNP_118828_190 | C/T | 0.234 | 2,030,017 | MTR_2g009690 |
| SNP_80231_27 | C/T | 0.234 | 2,163,048 | MTR_2g009940 |
| SNP_364434_97 | A/T | 0.359 | 2,163,084 | |

Genotype refers to the reference and alternative alleles; FDR is the false discovery rate obtained in an association test with the V+C content of 20 Vicia faba lines; Position is the position in bp on the Medicago truncatula chromosome 2.

3.2.3. Systematic Identification of Regulatory SNPs Associated with the V+C Content of *Vicia faba*

Following the studies of Xu et al. and Fu et al. in [67,68], we scanned the flanking sequences of the SNPs by applying the MATCH$^{TM}$ program [64] to systematically identify the SNPs that are likely to affect the binding affinity of transcription factors (TFs) and, thus, influence the gene expression level. This search was done for the 1976 SNPs that were located in the predicted *Vicia faba* promoters and which could be successfully mapped onto the *Medicago truncatula* genome. We considered results of this run with an MSS score $\geq$0.85 as putative TFBSs (as suggested in [69]). 9444 putative TFBSs were identified. SNPs that were located in putative TFBSs were considered as rSNPs. Their consequence types were determined by examining their predicted effects on the binding affinities of the TFs. The rSNPs, their consequence and related TFs with corresponding PWM names are given in Supplementary Table S4. The analysis of the 14 SNPs presented in Table 4 reveals that the binding affinities of 44 TFs to their 79 TFBSs were affected. Focusing on the two highly significant SNPs (SNP_341016_236 and SNP_341016_239) in the same promoter, we found that a nucleotide substitution in SNP_341016_236 is likely to entail severe consequences regarding TF binding affinity, namely loss and gain of TFBSs.

The substitution in SNP_341016_239 results in only a moderate change of the binding affinities of TFs (see Table 5). The remarkably different consequences of both SNPs indicate their considerably different influence for the precise and effective regulation of the corresponding gene, although their *p*-values are the same.

**Table 5.** The two SNPs with the strongest association to the V+C content and their consequences. The column **Allele** indicates for which allele of the SNP the binding site was found. **TFBS** refers to the name of the binding sites, which were named after their PWMs. The structure of the PWM names is given as: P$TFname_version, where "P$" stands for the PWMs used for the prediction of the TFBSs of plant TFs. "TFname" refers to the name of the transcription factor, and "_version" refers to the version of the PWM.

| SNP_ID | Allele | TFBS | MSS | Consequence |
|---|---|---|---|---|
| SNP_341016_236 | Ref | P$MYB4_01 | 0.945 | Loss of TFBS |
| SNP_341016_236 | Ref | P$MYB61_01 | 0.880 | Loss of TFBS |
| SNP_341016_236 | Alt | P$SQUA_01 | 0.870 | Gain of TFBS |
| SNP_341016_239 | Ref | P$MYB61_01 | 0.880 | Score change |
| SNP_341016_239 | Alt | P$MYB61_01 | 0.881 | Score change |

*3.3. Functional Analysis of the Candidate Gene and Transcription Factors*

The *Medicago truncatula* gene MTR_2g008620 is the gene which is located closest to the two highly significant SNPs. It is a beta-hydroxyacyl-ACP-dehydratase that is involved in the elongation of fatty acids as well as in the related metabolism of biotin [70,71]. A direct association with the creation of V+C which has been linked to the orotic acid pathway [72] is not obvious. This seems plausible since *Medicago truncatula* does not synthesize V+C. Of more interest are the transcription factors for which we found putative binding sites that are affected by the two SNPs. The TF SQUA belongs to the MADS-box domain group whose genes play vital roles in multiple aspects of plant development (for instance development of flowers, fruits and roots as well as regulating flowering time) [73,74]. Such genes regulate, for example, stem growth and early flowering in soybean [75] or vernalization response in wheat [76]. SQUA itself is involved in the determination of floral meristem and organ identity [77,78]. The MYB domain group is one of the largest families of TFs in plants. Its members are involved in the regulation of development, metabolism, the circadian rhythm, and responses to biotic and abiotic stresses in plants [74,79,80]. In *Medicago truncatula* multiple MYB TFs including MYB4 and MYB61 are involved in flavonoid biosynthesis during macrosclereid cell development [81]. MYB4 in particular regulates abiotic stress responses towards UV-B light and cadmium toxicity in *Arabidopsis*

*thaliana* [82,83] and cold in *Oryza sativa* [84]. It has also been shown to influence the biosynthesis of flavonoids [85]. MYB61 participates in the response to cold stress in *Medicago truncatula* [86]. In *Arabidopsis thaliana*, this TF is expressed in sink tissues, such as xylem, roots and developing seeds, and controls resource allocation influencing growth and development of the plant [87]. It has also been shown to affect trichome initiation, root development and stomatal aperture and it is necessary for the biosynthesis of gibberellin [88,89]. Furthermore, it is required for the seed coat mucilage deposition during the development of the seed coat epidermis [90,91]. This is a promising result considering that the seed coat is the suggested site of biosynthesis of the V+C compounds [92].

## 4. Conclusions

With their anti-nutritive effects the high vicine and convicine content has so far restricted the usage of *Vicia faba* as feed for livestock or as crop for human consumption. Identifying causal markers and understanding the mechanisms of regulation of the V+C content are important steps for breeding new cultivars with lower V+C content. This task is even more challenging since a complete and annotated reference genome for *Vicia faba* is still missing. In this work we harnessed the knowledge about regulatory regions in related species to train a convolutional neural network, which allowed us to predict those regions in *Vicia faba*. This model permitted us to classify DNA sequences as promoter or non-promoter without undertaking the considerable effort of assembling and annotating a reference genome. We applied this model to GBS data of *Vicia faba* for the identification of its putative promoter regions as well as regulatory SNPs therein. Our results show that we were able to detect two rSNPs significantly associated with the V+C production in *Vicia faba*. We suggest these rSNPs as promising candidates for marker-assisted selection. In particular, the associated transcription factor MYB61 could provide new insights into the molecular mechanisms underlying V+C. To the best of our knowledge, this is the first study which uses the gene annotations of related species of the *Leguminosae* family to predict promoters and rSNPs of *Vicia faba* based on the GBS data. The analysis approach that we presented here could potentially also be applied to other species that lack a reference genome which is still the case for many crop species.

**Supplementary Materials:** The following are available online at http://www.mdpi.com/2073-4425/11/6/614/s1, Table S1: Sensitivity values of the interspecies promoter classification using the intraspecies trained CNNs. Table S2: Specificity values of the interspecies promoter classification using the intraspecies trained CNNs. Table S3: MCC values of the interspecies promoter classification using the intraspecies trained CNNs. Table S4: Results of the TFBS analysis. Table S5: Overview of the 20 *Vicia faba* lines used in the analysis. File S6: Promoter sequences used for training the model. File S7: Non-promoter sequences used for training the model. Table S8: Alleles of the 20 *Vicia faba* lines for the two significantly associated rSNPs.

**Author Contributions:** M.G. designed and supervised the research. F.H. participated in the design of the study, prepared the data sets, conducted the bioinformatics analysis and developed the model together with M.W., P.P.D., and M.G. F.H. and M.K. performed the functional analysis of gene and transcription factors. A.O.S. and W.L. secured the funding for data acquisition. W.L. provided seed of the inbred lines, expertise with the crop plant and contributed to the training strategy. F.H., M.W., M.K., M.G. and A.O.S. interpreted the results and wrote the final version of the manuscript. M.G. and A.O.S. supervised the writing of the manuscript, conceived as well as managed the project. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by the Lower Saxony Ministry of Science and Culture, grant number MWK 11-76251-99-30/16.

**Acknowledgments:** We acknowledge support by the German Research Foundation and the Open Access Publication Funds of the University of Göttingen. We are grateful to Rebecca Tacke, Thomas Lange, and Wolfgang Ecke for providing valuable advice on some biological aspects. We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

**Conflicts of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

| | |
|---|---|
| SNP | Single Nucleotide Polymorphism |
| rSNP | Regulatory Single Nucleotide Polymorphism |
| CNN | Convolutional Neural Network |
| TF | Transcription Factor |
| TFBS | Transcription Factor Binding Site |
| V+C | Vicin and Convicine |
| FDR | False Discovery Rate |
| HMI | Horizontal Mutual Information |
| GTE | Generalized Topological Entropy |
| CPE | Core Promoter Element |
| PWM | Position Weight Matrix |
| MSS | Matrix Similarity Score |

## References

1. Deschamps, S.; Llaca, V.; May, G.D. Genotyping-by-Sequencing in Plants. *Biology* **2012**, *1*, 460–483. [CrossRef] [PubMed]
2. Muktar, M.S.; Teshome, A.; Hanson, J.; Negawo, A.T.; Habte, E.; Entfellner, J.D.; Lee, K.; Jones, C.S. Genotyping by sequencing provides new insights into the diversity of Napier grass (Cenchrus purpureus) and reveals variation in genome-wide LD patterns between collections. *Sci. Rep.* **2019**, *9*, 6936. [CrossRef] [PubMed]
3. Raman, H.; Raman, R.; Nelson, M.N.; Aslam, M.N.; Rajasekaran, R.; Wratten, N.; Cowling, W.A.; Kilian, A.; Sharpe, A.G.; Schondelmaier, J. Diversity array technology markers: genetic diversity analyses and linkage map construction in rapeseed (Brassica napus L.). *DNA Res.* **2011**, *19*, 51–65. [CrossRef] [PubMed]
4. Wenzl, P.; Raman, H.; Wang, J.; Zhou, M.; Huttner, E.; Kilian, A. A DArT platform for quantitative bulked segregant analysis. *BMC Genom.* **2007**, *8*, 196. [CrossRef]
5. He, J.; Zhao, X.; Laroche, A.; Lu, Z.; Liu, H.; Li, Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. *Front. Plant Sci.* **2014**, *5*, 484. [CrossRef]
6. Nguyen, N.H.; Premachandra, H.K.A.; Kilian, A.; Knibb, W. Genomic prediction using DArT-Seq technology for yellowtail kingfish Seriola lalandi. *BMC Genom.* **2018**, *19*, 107. [CrossRef]
7. Von Mark, V.C.; Kilian, A.; Dierig, D.A. Development of DArT marker platforms and genetic diversity assessment of the US collection of the new oilseed crop lesquerella and related species. *PLoS ONE* **2013**, *8*, e64062.
8. Morris, G.P.; Ramu, P.; Deshpande, S.P.; Hash, C.T.; Shah, T.; Upadhyaya, H.D.; Riera-Lizarazu, O.; Brown, P.J.; Acharya, C.B.; Mitchell, S.E.; et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 453–458. [CrossRef]
9. International Cassava Genetic Map Consortium. High-resolution linkage map and chromosome-scale genome assembly for cassava (Manihot esculenta Crantz) from 10 populations. *G3 Genes Genomes Genet.* **2015**, *5*, 133–144.
10. Soto, J.C.; Ortiz, J.F.; Perlaza-Jiménez, L.; Vásquez, A.X.; Lopez-Lavalle, L.A.B.; Mathew, B.; Léon, J.; Bernal, A.J.; Ballvora, A.; López, C.E. A genetic map of cassava (Manihot esculenta Crantz) with integrated physical mapping of immunity-related genes. *BMC Genom.* **2015**, *16*, 190. [CrossRef]
11. Elshire, R.J.; Glaubitz, J.C.; Sun, Q.; Poland, J.A.; Kawamoto, K.; Buckler, E.S.; Mitchell, S.E. A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. *PLoS ONE* **2011**, *6*, e19379. [CrossRef] [PubMed]
12. Cooper, J.W.; Wilson, M.H.; Derks, M.F.L.; Smit, S.; Kunert, K.J.; Cullis, C.; Foyer, C.H. Enhancing faba bean (Vicia faba L.) genome resources. *J. Exp. Bot.* **2017**, *68*, 1941–1953. [CrossRef] [PubMed]
13. Köpke, U.; Nemecek, T. Ecological services of faba bean. *Field Crop. Res.* **2010**, *115*, 217–233. [CrossRef]

14. Khazaei, H.; Purves, R.W.; Hughes, J.; Link, W.; O'Sullivan, D.M.; Schulman, A.H.; Björnsdotter, E.; Geu-Flores, F.; Nadzieja, M.; Andersen, S.U.; et al. Eliminating vicine and convicine, the main anti-nutritional factors restricting faba bean usage. *Trends Food Sci. Technol.* **2019**, *91*, 549–556. [CrossRef]
15. Arese, P.; Gallo, V.; Pantaleo, A.; Turrini, F. Life and Death of Glucose-6-Phosphate Dehydrogenase (G6PD) Deficient Erythrocytes - Role of Redox Stress and Band 3 Modifications. *Transfus. Med. Hemotherapy* **2012**, *39*, 328–334. [CrossRef] [PubMed]
16. Duc, G.; Sixdenier, G.; Lila, M.; Furstoss, V. Search of Genetic Variability for Vicine and Convicine Content in *Vicia faba* L.: A First Report of a Gene Which Codes for Nearly Zero-Vicine and Zero-Convicine Contents. In *Recent Advances of Research in Antinutritional Factors in Legume Seeds*; Huisman, J., van der Poel, A.F.B., Liener, I.E., Eds.; Wageningen Academic Publishers: Wageningen, The Netherlands, 1989; pp. 305–313.
17. Fang, L.; Ahn, J.K.; Wodziak, D.; Sibley, E. The human lactase persistence-associated SNP -13910*T enables in vivo functional persistence of lactase promoter-reporter transgene expression. *Hum. Genet.* **2012**, *131*, 1153–1159. [CrossRef]
18. De Gobbi, M.; Viprakasit, V.; Hughes, J.R.; Fisher, C.; Buckle, V.J.; Ayyub, H.; Gibbons, R.J.; Vernimmen, D.; Yoshinaga, Y.; de Jong, P.; et al. A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* **2006**, *312*, 1215–1217. [CrossRef] [PubMed]
19. Ordovás, L.; Roy, R.; Pampín, S.; Zaragoza, P.; Osta, R.; Rodríguez-Rey, J.C.; Rodellar, C. The g.763G>C SNP of the bovine FASN gene affects its promoter activity via Sp-mediated regulation: implications for the bovine lactating mammary gland. *Physiol. Genom.* **2008**, *34*, 144–148. [CrossRef] [PubMed]
20. Ryan, M.T.; Hamill, R.M.; O'Halloran, A.M.; Davey, G.C.; McBryan, J.; Mullen, A.M.; McGee, C.; Gispert, M.; Southwood, O.I.; Sweeney, T. SNP variation in the promoter of the PRKAG3 gene and association with meat quality traits in pig. *BMC Genet.* **2012**, *13*, 66. [CrossRef]
21. Barkova, O.Y.; Sazanova, K.A.; Fomichev, K.A.; Malewski, T.; Parada, R.; Kawka, M.; Jaszczak, K.; Sazanov, A.A. Associations of new rSNPs with eggshell thickness in Rhode Island layers. *Anim. Sci. Pap. Rep.* **2013**, *31*, 165–172.
22. Konishi, S.; Izawa, T.; Lin, S.Y.; Ebana, K.; Fukuta, Y.; Sasaki, T.; Yano, M. An SNP caused loss of seed shattering during rice domestication. *Science* **2006**, *312*, 1392–1396. [CrossRef]
23. Fickett, J.W.; Hatzigeorgiou, A.G. Eukaryotic Promoter Recognition. *Genome Res.* **1997**, *7*, 861–878. [CrossRef]
24. Shahmuradov, I.A.; Solovyev, V.V.; Gammerman, A.J. Plant promoter prediction with confidence estimation. *Nucleic Acids Res.* **2005**, *33*, 1069–1076. [CrossRef]
25. Ohler, U. Identification of core promoter modules in Drosophila and their application in accurate transcription start site prediction. *Nucleic Acids Res.* **2006**, *34*, 5943–5950. [CrossRef]
26. Morey, C.; Mookherjee, S.; Rajasekaran, G.; Bansal, M. DNA Free Energy-Based Promoter Prediction and Comparative Analysis of Arabidopsis and Rice Genomes. *Plant Physiol.* **2011**, *156*, 1300–1315. [CrossRef]
27. Azad, A.K.M.; Shahid, S.; Noman, N.; Lee, H. Prediction of plant promoters based on hexamers and random triplet pair analysis. *Algorithms Mol. Biol.* **2011**, *6*, 19. [CrossRef]
28. Lai, H.; Zhang, Z.; Su, Z.; Su, W.; Ding, H.; Chen, W.; Lin, H. iProEP: A Computational Predictor for Predicting Promoter. *Mol. Ther. Nucleic Acids* **2019**, *17*, 337–346. [CrossRef]
29. Abeel, T.; Saeys, Y.; Rouzé, P.; Van de Peer, Y. ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics* **2008**, *24*, i24–i31. [CrossRef]
30. Anwar, F.; Baker, S.M.; Jabid, T.; Mehedi Hasan, M.; Shoyaib, M.; Khan, H.; Walshe, R. Pol II promoter prediction using characteristic 4-mer motifs: A machine learning approach. *BMC Bioinform.* **2008**, *9*, 414. [CrossRef]
31. Umarov, R.K.; Solovyev, V.V. Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. *PLoS ONE* **2017**, *12*, e171410. [CrossRef]
32. Umarov, R.; Kuwahara, H.; Li, Y.; Gao, X.; Solovyev, V. Promoter analysis and prediction in the human genome using sequence-based deep learning models. *Bioinformatics* **2019**, *35*, 2730–2737. [CrossRef]
33. Triska, M.; Solovyev, V.; Baranova, A.; Kel, A.; Tatarinova, T.V. Nucleotide patterns aiding in prediction of eukaryotic promoters. *PLoS ONE* **2017**, *12*, 1–28. [CrossRef]
34. Qian, Y.; Zhang, Y.; Guo, B.; Ye, S.; Wu, Y.; Zhang, J. An Improved Promoter Recognition Model Using Convolutional Neural Network. In Proceedings of the 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 23–27 July 2018; Volume 1, pp. 471–476. [CrossRef]

35. Oubounyt, M.; Louadi, Z.; Tayara, H.; Chong, K.T. DeePromoter: Robust Promoter Predictor Using Deep Learning. *Front. Genet.* **2019**, *10*, 286. [CrossRef]
36. Pachganov, S.; Murtazalieva, K.; Zarubin, A.; Sokolov, D.; Chartier, D.R.; Tatarinova, T.V. TransPrise: A novel machine learning approach for eukaryotic promoter prediction. *PeerJ* **2019**, *7*, e7990. [CrossRef]
37. Kumari, S.; Ware, D. Genome-Wide Computational Prediction and Analysis of Core Promoter Elements across Plant Monocots and Dicots. *PLoS ONE* **2013**, *8*, e79011. [CrossRef]
38. Shahmuradov, I.A.; Umarov, R.K.; Solovyev, V.V. TSSPlant: a new tool for prediction of plant Pol II promoters. *Nucleic Acids Res.* **2017**, *45*, e65. [CrossRef]
39. Goubert, C.; Modolo, L.; Vieira, C.; ValienteMoro, C.; Mavingui, P.; Boulesteix, M. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow Fever Mosquito (Aedes aegypti). *Genome Biol. Evol.* **2015**, *7*, 1192–1205. [CrossRef]
40. Yuan, S.; Xia, Y.; Zheng, Y.; Zeng, X. Next-generation sequencing of mixed genomic DNA allows efficient assembly of rearranged mitochondrial genomes in *Amolops chunganensis* and *Quasipaa boulengeri*. *PeerJ* **2016**, *4*, e2786. [CrossRef]
41. Grabherr, M.G.; Haas, B.J.; Yassour, M.; Levin, J.Z.; Thompson, D.A.; Amit, I.; Adiconis, X.; Fan, L.; Raychowdhury, R.; Zeng, Q.; et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **2011**, *29*, 644–652. [CrossRef]
42. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef]
43. Hwang, S.; Kim, E.; Lee, I.; Marcotte, E.M. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Sci. Rep.* **2015**, *5*, 17875. [CrossRef] [PubMed]
44. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359. [CrossRef]
45. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079. [CrossRef] [PubMed]
46. Chang, C.C.; Chow, C.C.; Tellier, L.CAM.; Vattikuti, S.; Purcell, S.M.; Lee, J.J. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **2015**, *4*. [CrossRef]
47. Howe, K.L.; Contreras-Moreira, B.; De Silva, N.; Maslen, G.; Akanni, W.; Allen, J.; Alvarez-Jarreta, J.; Barba, M.; Bolser, D.M.; Cambell, L.; et al. Ensembl Genomes 2020—Enabling non-vertebrate genomic research. *Nucleic Acids Res.* **2019**, gkz890. [CrossRef] [PubMed]
48. Kinsella, R.J.; Kähäri, A.; Haider, S.; Zamora, J.; Proctor, G.; Spudich, G.; Almeida-King, J.; Staines, D.; Derwent, P.; Kerhornou, A.; et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011**, *2011*, bar030. [CrossRef]
49. Humann, J.L.; Jung, S.; Cheng, C-H.; Lee, T.; Zheng, P.; Frank, M.; McGaughey, D.; Scott, K.; Buble, K.; Yu, J.; et al. Cool Season Food Legume Genome Database: A resource for pea, lentil, faba bean and chickpea genetics, genomics and breeding. In Proceedings of the International Plant and Animal Genome Conference, San Diego, CA, USA, 12–16 January 2019.
50. Lichtenstein, F.; Antoneli, F.; Briones, M.R.S. MIA: Mutual Information Analyzer, a graphic user interface program that calculates entropy, vertical and horizontal mutual information of molecular sequence sets. *BMC Bioinform.* **2015**, *16*, 409. [CrossRef]
51. Schmitt, A.O.; Herzel, H. Estimating the entropy of DNA sequences. *J. Theor. Biol.* **1997**, *188*, 369–377. [CrossRef]
52. Jin, S.; Tan, R.; Jiang, Q.; Xu, L.; Peng, J.; Wang, Y.; Wang, Y. A Generalized Topological Entropy for Analyzing the Complexity of DNA Sequences. *PLoS ONE* **2014**, *9*, e88519. [CrossRef]
53. Li, J.; Zhang, L.; Li, H.; Ping, Y.; Xu, Q.; Wang, R.; Tan, R.; Wang, Z.; Liu, B.; Wang, Y. Integrated entropy-based approach for analyzing exons and introns in DNA sequences. *BMC Bioinform.* **2019**, *20*, 283. [CrossRef]
54. Al-Ajlan, A.; El Allali, A. CNN-MGP: Convolutional neural networks for metagenomics gene prediction. *Interdiscip. Sci. Comput. Life Sci.* **2019**, *11*, 628–635. [CrossRef] [PubMed]
55. Chollet, F.; Allaire, J.J. *Deep Learning with R*; Manning Publications: Shelter Island, NY, USA, 2018.

56. Li, Q.; Cai, W.; Wang, X.; Zhou, Y.; Feng, D.D.; Chen, M. Medical image classification with convolutional neural network. In Proceedings of the 2014 13th International Conference on Control Automation Robotics & Vision (ICARCV), Singapore, 10–12 December 2014; pp. 844–848.
57. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef] [PubMed]
58. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
59. Kingma, D.P; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:cs.LG/1412.6980.
60. Yu, K.; Xu, W.; Gong, Y. Deep learning with kernel regularization for visual recognition. In *Advances in Neural Information Processing Systems*; Koller, D., Schuurmans, D., Bengio, Y., Bottou, L., Eds.; Curran Associates Inc.: Red Hook, NY, USA, 2009; pp. 1889–1896.
61. Chollet, F. Keras. 2015. Available online: https://keras.io (accessed on 28 May 2020).
62. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. Available online: tensorflow.org (accessed on 28 May 2020).
63. Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS ONE* **2017**, *12*, e0177678. [CrossRef]
64. Kel, A.E.; Gößling, E.; Cheremushkin, E.; Kel-Margoulis, O.V.; Wingender, E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**, *31*, 3576–3579. [CrossRef]
65. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* **2008**, *9*, 326–332. [CrossRef]
66. Camacho, C.; Coulouris, G.A.V.M.N.P.J.B.K.M.T. BLAST+: architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [CrossRef]
67. Xu, Z.; Taylor, J.A. SNPinfo: Integrating GWAS and candidate gene information into functional SNP selection for genetic association studies. *Nucleic Acids Res.* **2009**, *37*, W600–W605. [CrossRef]
68. Fu, Y.; Liu, Z.; Lou, S.; Bedford, J.; Mu, X.J.; Yip, K.Y.; Khurana, E.; Gerstein, M. FunSeq2: A framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **2014**, *15*, 480. [CrossRef]
69. Gearing, L.J.; Cumming, H.E.; Chapman, R.; Finkel, A.M.; Woodhouse, I.B.; Luu, K.; Gould, J.A.; Forster, S.C.; Hertzog, P.J. CiiiDER: A tool for predicting and analysing transcription factor binding sites. *PLoS ONE* **2019**, *14*, e0215495. [CrossRef]
70. Heath, R.J.; Rock, C.O. Roles of the FabA and FabZ $\beta$-Hydroxyacyl-Acyl Carrier Protein Dehydratases in Escherichia coli Fatty Acid Biosynthesis. *J. Biol. Chem.* **1996**, *271*, 27795–27801. [CrossRef] [PubMed]
71. Lin, S.; Hanson, R.E.; Cronan, J.E. Biotin synthesis begins by hijacking the fatty acid synthetic pathway. *Nat. Chem. Biol.* **2010**, *6*, 682–688. [CrossRef] [PubMed]
72. Brown, E.G.; Roberts, F.M. Formation of vicine and convicine by *Vicia faba*. *Phytochemistry* **1972**, *11*, 3203–3206. [CrossRef]
73. Smaczniak, C.; Immink, R.G.H.; Angenent, G.C.; Kaufmann, K. Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. *Development* **2012**, *139*, 3081–3098. [CrossRef]
74. Riechmann, J.L.; Ratcliffe, O.J. A genomic perspective on plant transcription factors. *Curr. Opin. Plant Biol.* **2000**, *3*, 423–434. [CrossRef]
75. Ping, J.; Liu, Y.; Sun, L.; Zhao, M.; Li, Y.; She, M.; Sui, Y.; Lin, F.; Liu, X.; Tang, Z.; et al. Dt2 Is a Gain-of-Function MADS-Domain Factor Gene That Specifies Semideterminacy in Soybean. *Plant Cell* **2014**, *26*, 2831–2842. [CrossRef]
76. Danyluk, J.; Kane, N.A.; Breton, G.; Limin, A.E.; Fowler, D.B.; Sarhan, F. TaVRT-1, a Putative Transcription Factor Associated with Vegetative to Reproductive Transition in Cereals. *Plant Physiol.* **2003**, *132*, 1849–1860. [CrossRef]
77. West, A.G.; Sharrocks, A.D.; Causier, B.E.; Davies, B. DNA binding and dimerisation determinants of Antirrhinum majus MADS-box transcription factors. *Nucleic Acids Res.* **1998**, *26*, 5277–5287. [CrossRef]
78. Theißen, G.; Melzer, R.; Rümpler, F. MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. *Development* **2016**, *143*, 3259–3271. [CrossRef]
79. Dubos, C.; Stracke, R.; Grotewold, E.; Weisshaar, B.; Martin, C.; Lepiniec, L. MYB transcription factors in *Arabidopsis*. *Trends Plant Sci.* **2010**, *15*, 573–581. [CrossRef]

80. Roy, S. Function of MYB domain transcription factors in abiotic stress and epigenetic control of stress response in plant genome. *Plant Signal. Behav.* **2016**, *11*, e1117723. [CrossRef] [PubMed]
81. Fu, F.; Zhang, W.; Li, Y.; Wang, H.L. Establishment of the model system between phytochemicals and gene expression profiles in Macrosclereid cells of Medicago truncatula. *Sci. Rep.* **2017**, *7*, 2580. [CrossRef] [PubMed]
82. Jin, H.; Cominelli, E.; Bailey, P.; Parr, A.; Mehrtens, F.; Jones, J.; Tonelli, C.; Weisshaar, B.; Martin, C. Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. *EMBO J.* **2000**, *19*, 6150–6161. [CrossRef] [PubMed]
83. Agarwal, P.; Banerjee, S.; Mitra, M.; Roy, S. MYB4 transcription factor, A member of R2R3-type MYB family protein regulates Cd tolerance via activation of antioxidant defense and glutathione (GSH) dependent pathway in Arabidopsis thaliana. In Proceedings of the XIV International Geographical Union (IGU)-India Conference, Burdwan, India, 6–8 March 2020.
84. Vannini, C.; Locatelli, F.; Bracale, M.; Magnani, E.; Marsoni, M.; Osnato, M.; Mattana, M.; Baldoni, E.; Coraggio, I. Overexpression of the rice Osmyb4 gene increases chilling and freezing tolerance of Arabidopsis thaliana plants. *Plant J.* **2004**, *37*, 115–127. [CrossRef] [PubMed]
85. Wang, X.; Wu, J.; Guan, M.; Zhao, C.; Geng, P.; Zhao, Q. Arabidopsis MYB4 plays dual roles in flavonoid biosynthesis. *Plant J.* **2020**, *101*, 637–652. [CrossRef]
86. Zhang, Z.; Hu, X.; Zhang, Y.; Miao, Z.; Xie, C.; Meng, X.; Deng, J.; Wen, J.; Mysore, K.S.; Frugier, F.; et al. Opposing Control by Transcription Factors MYB61 and MYB3 Increases Freezing Tolerance by Relieving C-Repeat Binding Factor Suppression. *Plant Physiol.* **2016**, *172*, 1306–1323. [CrossRef]
87. Romano, J.M.; Dubos, C.; Prouse, M.B.; Wilkins, O.; Hong, H.; Poole, M.; Kang, K.; Li, E.; Douglas, C.J.; Western, T.L.; et al. AtMYB61, an R2R3-MYB transcription factor, functions as a pleiotropic regulator via a small gene network. *New Phytol.* **2012**, *195*, 774–786. [CrossRef]
88. Matías-Hernández, L.; Jiang, W.; Yang, K.; Tang, K.; Brodelius, P.E.; Pelaz, S. AaMYB1 and its orthologue AtMYB61 affect terpene metabolism and trichome development in Artemisia annua and Arabidopsis thaliana. *Plant J.* **2017**, *90*, 520–534. [CrossRef]
89. Liang, Y.; Dubos, C.; Dodd, I.C.; Holroyd, G.H.; Hetherington, A.M.; Campbell, M.M. AtMYB61, an R2R3-MYB Transcription Factor Controlling Stomatal Aperture in Arabidopsis thaliana. *Curr. Biol.* **2005**, *15*, 1201–1206. [CrossRef]
90. Arsovski, A.A.; Villota, M.M.; Rowland, O.; Subramaniam, R.; Western, T.L. MUM ENHANCERS are important for seed coat mucilage production and mucilage secretory cell differentiation in Arabidopsis thaliana. *J. Exp. Bot.* **2009**, *60*, 2601–2612. [CrossRef]
91. Penfield, S.; Meissner, R.C.; Shoue, D.A.; Carpita, N.C.; Bevan, M.W. MYB61 Is Required for Mucilage Deposition and Extrusion in the Arabidopsis Seed Coat. *Plant Cell* **2001**, *13*, 2777–2791. [CrossRef] [PubMed]
92. Ramsay, G.; Griffiths, D.W. Accumulation of vicine and convicine in *Vicia faba* and *V. narbonensis*. *Phytochemistry* **1996**, *42*, 63–67. [CrossRef]