agriculture

MDPI

Review

Advances in Integrating Genomics and Bioinformatics in the Plant Breeding Pipeline

Haifei Hu, Armin Scheben and David Edwards *

School of Biological Sciences and Institute of Agriculture, The University of Western Australia, Perth 6009, Australia; haifei.hu@research.uwa.edu.au (H.H.); armin.scheben@research.uwa.edu.au (A.S.)

* Correspondence: dave.edwards@uwa.edu.au; Tel.: +61-8-6488-2415

Received: 9 May 2018; Accepted: 28 May 2018; Published: 31 May 2018

Abstract: With the global human population growing rapidly, agricultural production must increase to meet crop demand. Improving crops through breeding is a sustainable approach to increase yield and yield stability without intensifying the use of fertilisers and pesticides. Current advances in genomics and bioinformatics provide opportunities for accelerating crop improvement. The rise of third generation sequencing technologies is helping overcome challenges in plant genome assembly caused by polyploidy and frequent repetitive elements. As a result, high-quality crop reference genomes are increasingly available, benefitting downstream analyses such as variant calling and association mapping that identify breeding targets in the genome. Machine learning also helps identify genomic regions of agronomic value by facilitating functional annotation of genomes and enabling real-time high-throughput phenotyping of agronomic traits in the glasshouse and in the field. Furthermore, crop databases that integrate the growing volume of genotype and phenotype data provide a valuable resource for breeders and an opportunity for data mining approaches to uncover novel trait-associated candidate genes. As knowledge of crop genetics expands, genomic selection and genome editing hold promise for breeding diseases-resistant and stress-tolerant crops with high yields.

Keywords: breeding; crops; genomics; third generation sequencing

1. Introduction

Humans depend on crops for over two-thirds of their daily energy intake [1,2]. As the global human population grows, agriculture (crop cultivation) is under increasing pressure to produce higher crop yields [3]. Additionally, climate change, limited availability of land and water shortages are posing further agricultural challenges [4]. To increase crop yields while reducing the environmental impact of agriculture, genomics is accelerating crop breeding by helping systematically leverage the genetic components of agronomic traits [5,6]. Crop genome sequences provide an important foundation for identifying agronomically relevant variation. During the last decade, the decreasing cost of DNA sequencing has led to a rapid rise in the size of crop genomic data, which represents a substantial opportunity for breeders [7].

Although plant genome assembly (generating a genome sequence from fragmented sequencing reads) is still hampered by frequent long repetitive regions, large genome sizes and frequent polyploidy, advances in sequencing technologies and bioinformatics tools have allowed rapid progress since the sequencing and assembly of the rice genome in 2005 [8].

Although rice was still sequenced using bacterial artificial chromosomes (BAC) and Sanger sequencing, the grape genome published in 2007 was the first to use a combination of the less costly 454 sequencing and Sanger sequencing [9]. Two years later, Illumina short reads were combined with Sanger sequencing to assemble the cucumber genome [10], marking the start of the rapid adoption of

next generation sequencing (NGS) [11]. By 2013, 55 plant genomes had been sequenced, with 40 of these belonging to crops [12]. Third generation sequencing technologies capable of generating long reads greater than 10 kb in length were developed in recent years, providing a further useful tool for crop genome sequencing. Today, there are over 260 land plant nuclear genomes publicly available in GenBank, including most major crops.

Crop breeding has long relied on cycles of phenotypic selection and crossing, which generate superior genotypes through genetic recombination. When genome sequences are available, all genes and genetic variants contributing to agronomics traits can be identified and changes made during breeding processes can be assessed at the genotype level. Because of the ready availability of genomic data for breeders today, genomics plays an increasingly important role in all aspects of crop breeding, such as quantitative trait loci (QTL) mapping and genome-wide association studies (GWAS), where genomic sequencing of crop populations can allow gene-level resolution of agronomic variation. For example, advances in genomics-based breeding allow the identification of genetic variation in crop species, which can be applied to produce climate resilient crops [5,13,14]. In a different approach, genomic selection (GS) harnesses genome-wide genetic variants to avoid the need for repeated phenotyping in breeding cycles.

Bioinformatics is crucial for processing and analysing large genomic datasets and gaining functional insights into plant genomes [15,16]. Although genome assembly, sequence alignment and variant calling are standard bioinformatics tasks when analysing sequencing data, the algorithms required are non-trivial and there are different competing computational approaches with unique biases [17–19]. For alignment of third generation sequencing data, many tools developed for short reads perform poorly when aligning long reads [20]. A challenge for assembly and alignment tools is combining different data types such as short reads and long reads to reduce the impact of unique biases. Downstream analyses such as comparative genomic analysis, variant calling and GWAS can provide comprehensive information to facilitate crop improvement [21]. Variant callers differ in their ability to call indels and in their biases towards calling heterozygous and reference variants [19]. Software designed for carrying out GWAS may apply models of varying complexity to address the effects of population structure. Applying the right bioinformatics tool for the right application is therefore crucial. Although there is a wide range of tools available to mine genomes and variant data, processing the growing amounts of genomic data and selecting the appropriate analyses is also a leading challenge faced by researchers in crop genomics [22]. Adaptation of existing crop databases such as GrainGenes [23] and Gramene [24] as well as development of novel databases such as the Wheat Information System (WheatIS) [25] will help store the deluge of data and make it more accessible to breeders.

An interdisciplinary approach is needed for plant breeding in the 21st century to identify and resolve breeding challenges and improve crop production [26]. Together with novel glasshouse technologies to accelerate plant development [27], genomics and bioinformatics play an important role in increasing the production rate of improved crop cultivars. Nevertheless, the vast amounts of genotypic and phenotypic data available create an enormous challenge to integrate diverse data outputs for breeding [28]. Integrating phenotypes, genomics and bioinformatics tools and resources in public and private breeding pipelines will address this challenge and help deliver breeding targets [29]. In this review, we discuss advances in genomics and bioinformatics, suggesting how these can be integrated to allow precise breeding and overcome bottlenecks in crop improvement.

2. Third Generation Sequencing to Improve Crop Genome Assemblies

Since the introduction of NGS, genome sequencing and resequencing have become standard in many disciplines of plant biology [11]. However, there are important limitations of NGS, such as inherent biases and ambiguous alignment of repetitive elements, which leads to highly fragmented draft genome assemblies and complicates the study of hidden indels and structural variants [20]. The emergence of third generation sequencing, including Pacific Biosciences (PacBio) single-molecule

real-time sequencing and Oxford Nanopore Technologies (ONT) sequencing, has enabled the generation of long reads and allowed production of more accurate and contiguous genome assemblies [30–32]. Third generation sequencing helps generate high-quality whole genome *de novo* assemblies, using reads spanning complex regions such as those with high levels of repetitive sequence and shed light on the remaining complex of repeat sequences and other structural variants. Moreover, the full-length sequenced transcripts produced by third generation sequencing techniques (isoform sequencing) allow the precise study of exons, splice sites and alternatively spliced regions, which is useful in improving genome annotation [30]. A fully assembled and well-annotated genome will allow breeders to discover genes related to agronomic traits, determine their location and function as well as develop genome-wide molecular markers.

Combining long-read sequencing with long-range mapping technologies and chromosome conformation capture has put highly contiguous chromosome-level crop genome assemblies within grasp even for smaller laboratories and non-model crop species [33]. New optical mapping platforms such as BioNano Genomics allow rapid labelling of long DNA molecules over 250 kb, enabling detection of structural variants and generation of a high-quality scaffolding at low cost. For example, using PacBio sequencing and optical mapping from BioNano, the assembly of the desiccation-tolerant grass species *Oropetium thomaeum* achieves a contig N50 of 2.4 Mb with over 99.5% genome coverage [34], a contiguity as high as that of model plant genomes such as Arabidopsis (TAIR10), rice (V 7) and *Brachypodium distachyon* (V 2.1). Optical mapping data was also used to generate a high-resolution map of the wheat chromosome 7DS with contigs showing an N50 of 1.3 Mb [35]. Long-read sequencing also uncovers repetitive regions with high accuracy. In *O. thomaeum*, long reads helped identify 18 telomeric regions, nine centromeric satellites and 3247 intact long terminal repeats in 358 families [34]. These repetitive sequences could not have been captured by short reads, potentially resulting in a miscalculation of their content. Another third-generation mapping innovation is chromosome conformation capture sequencing (Hi-C) [36], which is based on naturally physical close ligation of DNA fragments. Integration of Hi-C data and optical mapping allow further improvement of chromosome phasing and scaffolding. By combining short reads and optical and chromatin interaction mapping data, Mascher et al. [37] assembled the highly repetitive and polyploid barley genome, attaining an N50 value of 1.9 Mb. The higher level of sequence contiguity provided by third generation sequencing can facilitate genomics-based breeding approaches including trait mapping. For example, using a long read assembly of Arabidopsis, it was possible to define the growth-related SG3i QTL region that could not be previously resolved with Sanger sequencing [38].

The most powerful application of third-generation sequencing for breeding is the assembly of improved highly contiguous crop genomes. When selecting a sequencing approach for a crop genome assembly project, it is important to consider the size of the genome, ploidy, levels of repetitive content and the available funds. The current choice is mainly between PacBio, ONT and NGS, which can be used in combination with each other and supplemented with other long-range technologies such as BioNano and HiC. While costs for sequencing vary substantially between providers and countries, third generation sequencing can remain an order of magnitude costlier than NGS (Table 1). When sufficient funding is available, deep sequencing (>30×) using PacBio is likely to yield the most accurate, contiguous crop genome because it is less error prone than ONT sequencing [38] and existing sequencing errors can often be self-corrected at deep coverages [20]. However, the cost of PacBio deep sequencing may be prohibitive, particularly for large genomes. At a cost of $900/Gb with the PacBio RS II (Table 1), 30× coverage of the average-sized 389 Mb rice genome [8] would cost over $10,000. For genomes with repeat regions >100 kb and high levels of heterozygosity, the ultra-long reads (>100 kb) generated by ONT sequencing offer a unique advantage of this technology as in this case ultra-long reads improve assembly quality most [20]. When costs limit the depth of third generation sequencing to <30×, short reads can make an important contribution to error correct long reads before assembly [39] and to polish completed assemblies by correcting single base errors and small indels [38]. For resequencing studies of crop populations, third generation sequencing is therefore generally not

cost-effective and NGS remains the preferred approach. As the cost of third generation sequencing costs decreases, the application of these methods may extend into a broader range of genomic analysis beyond genome assembly.

Table 1. Comparison of the three main sequencing technologies for crop genomics. Corrected error rates are calculated after applying error correction methods that do not require additional data. Prices are in US$.

	Illumina Short Reads (HiSeq)	**PacBio (RSII/Sequel)**	**Oxford Nanopore Technologies (MinION)**
Sequencing method	Sequencing by synthesis	Single-molecule	Single-molecule
Average read length	125 bp–300 bp	10–15 kb (up to 100 kb)	10–15 kb (up to 1Mb)
Raw error rate	~0.1%	10–15%	10–38%
Corrected error rate	-	~1%	1–9%
Corrected error rate with short read polish	-	>0.1%	0.1–2.8%
Cost/Gb with library prep *	$30–$45	$240–$900	$160–$250
Applications for crop breeding	Resequencing studies; error correction of long reads	De novo assembly; scaffolding; gap filling	De novo assembly of genomes with long repetitive regions and high heterozygosity, scaffolding; gap filling
References	[11,40–42]	[34,40,43,44]	[45–49]

* Costs are estimates and can vary substantially between sequencing providers.

3. Integrated Crop Databases

With third generation sequencing technologies and other 'omics' technologies emerging, large volumes of data are available to investigate crop traits from the gene level to the population level [50]. Although essential sequence repositories such as Genbank [51], European Molecular Biological Laboratory (EMBL) [52], PlantGDB [53] and Phytozome [54] play an essential role, they mainly focus on storing and managing genomic data without integrating variant or phenotype data from other sources. This makes it more challenging for breeders and plant biologists to link genotype to phenotype, which often requires data on genomics, epigenomics, phenotypes and environments. Although come crop databases integrating these data exist, for example, marker and expression data are integrated in GrainGenes [23], additional databases are needed to address this gap in major repositories [55].

The task of creating an integrative crop database by combining annotated genome sequences, gene functions, interaction networks and trait phenotypes is challenging as the relevant data are dispersed in numerous databases with various data formats and different quality and coverage. Intelligent mining of large-scale crop databases is required to merge complex data resources and allow gene discovery and crop improvement [56]. To integrate biological data from different resources, a web-based intelligent mining tool, KnetMiner (Knowledge Network Miner) has been developed to search for links and concepts in biological knowledge networks, which enables the discovery of novel connections between traits and genes [56]. There are four main steps in the KnetMiner approach: (1) integrating diverse biological data into a knowledge graph, (2) improving the knowledge graph with text-mining of the literature, (3) identifying the link between genes and evidence nodes, (4) applying the evidence-based gene ranking algorithm and visualising the integrated data. Currently, KnetMiner has been employed for constructing integrative databases for important crops such as barley and wheat, providing insights into indirect associations between distant traits and biological processes. In barley, a gene-evidence network was applied to infer a connection between the MLOC_10687.2 and seed width phenotype, which show great potential in barley production improvement [57]. Progress is underway in wheat and rice to further develop the single information systems available for these crops [25,58], allowing broad querying across integrated databases. By ongoing development of integrative crop database

with advances in data mining techniques, breeders can understand a complex trait better and identify trait-associated candidate genes, which is beneficial for crop improvement [59].

4. Applying Integrative Genomics to Trait Discovery and Crop Improvement

4.1. Mining Quantitative Trait Loci Studies

The analysis of QTL enables estimation of genetic regions linked to quantitative phenotypic traits, bridging the gap between genomics and the field [60]. However, with the increasing number of QTL studies being conducted and reported in plants, a new challenge to identify high-quality candidate loci and further improve crop breeding is to integrate information from different QTL studies. In this case, meta-analysis, a tool to pool the outcomes of a range of studies and predict the location of QTL more precisely than individual studies, is required to use existing resources fully [61]. Bioinformatics tools are available for efficiently carrying out meta-QTL analysis. For instance, using statistics and a consensus model, a computational package called MetaQTL can reduce the length of the confidence interval of QTL, leading to precise estimation of the correct QTL location and effect [62]. Other bioinformatics tools such as solQTL and RASQUAL further offer the convenience of low bias QTL analysis, visualizing the QTL data and linking the QTL data with other genome databases [63,64]. To date, meta-QTL analyses have been performed to map traits related to crop development and abiotic and biotic responses in maize [65], cotton [66], soybean [67] and wheat [68]. For example, meta-QTL analysis has been used to identify five groups of yield and yield-related candidate genes in wheat with 195 molecular markers and 197 ESTs reported from 55 wheat QTL studies in the last 14 years [68]. Moreover, 37 QTLs related to nitrogen use efficiency were identified in maize [65]. Similarly, 20 consensus QTLs and their related markers were narrowed down by meta-QTL from a combination of QTL studies in last 20 years, which provides a foundation for gene mining and crop improvement in soybean [67]. QTL mapping remains a powerful method to link an observed agronomic trait to a genomic region. It provides a high detection power for scanning the complete crop genome and identifies rare alleles using limited genetic markers. Because QTL mapping is hypothesis-driven, it is applied when a trait of interest segregating in a mapping population is available. However, QTL mapping suffers from two fundamentals shortcomings: (1) low resolution caused by coarse mapping making it hard to differentiate pleiotropic and physically adjacent genes [69], (2) only allelic diversity present in the parents of the segregating population can be assayed [70]. To overcome these limitations of QTL mapping, GWAS can be employed to pinpoint genomic regions linked to traits in diverse, unrelated populations.

4.2. Genome-Wide Association Studies for Identifying Breeding Targets

In contrast to QTL analysis conducted on bi-parental populations derived from controlled crosses, GWAS relies on natural populations, providing higher resolution to identify multiple recombination events and explore the natural variations associated with phenotypical differences. GWAS leverages linkage disequilibrium to detect links between genotype and phenotype in crop species, achieving higher mapping resolution than QTL analysis [71]. When the aim of the breeder is an exploratory analysis to identify a broad range of genomic leads, GWAS is preferred to QTL analysis. Association studies are more likely than QTL analysis to identify specific candidate genes that can be directly introgressed into crop germplasm to improve crops [72]. GWAS has been carried out in a variety of crops such as rice, soybean, maize, wheat and canola [73–76]. In *Oryza sativa indica*, 517 landraces were sequenced, identifying around 3.6 million single nucleotide polymorphisms (SNPs) [73]. Using a GWAS of 14 agronomic traits, over 36% of the phenotypic variance could be explained by the identified loci, allowing further discovery of trait-related genes and alleles for crop improvement. Through a GWAS of maize, Tian et al. [75] revealed the architecture of leaf traits and found that variation in *liguleless* genes can result in upright leaves.

Advances in bioinformatics tools offer extra opportunities for conducting GWAS studies. For instance, PLINK is a widely used bioinformatics tool for GWAS, employing a standard regression analysis to associate genotypes with phenotypes [77]. However, for rare variants, standard regression cannot provide sufficient sensitivity for GWAS analysis [78]. TASSEL is another common GWAS tool and implements a mixed linear model incorporating population and family structure in the analysis [79]. Unlike PLINK, TASSEL can thus control for population effects. Other enhanced GWAS bioinformatics tools such as GAPIT have also been developed to computationally efficiently handle a large dataset containing over 1 million SNPs within 10,000 individuals using the compressed mixed linear model and model-based prediction and selection method [80].

4.3. Forward and Reverse Genetic Screening

Forward genetic screening is a widely used breeding tool and identifies and characterizes genes based on a known phenotype [81]. Reverse genetic screening, on the other hand, determines the phenotypic effect of altered sequences of specific genes or regulatory regions [82]. The starting point of forward genetics, including most QTL analyses, is a target phenotype that segregates in a population. Reverse genetics, on the other hand, can be used to functionally characterize known genes identified by QTL analysis or GWAS, with a mutation panel or a transformed line as a starting point. The role played by both forward and reverse genetic screening is essential to identify functional variation associated with agronomic traits such as tolerance to abiotic and biotic stresses, disease resistance, increased yield and improved nutritional quality. Forward genetic screening can improve gene cloning and marker development. Rather than performing whole genome sequencing to identify many sequences that are not related to heritable phenotypes, exome sequencing used in forward genetic screening allows selective screening of coding regions, excluding intergenic sequences. In rice, it was shown that to recover induced mutations it is sufficient to sequence 20 Mb of the 389 Mb genome [83]. Reverse genetic screening has been applied to crop functional genomics and breeding. Targeted Induced Local Lesions IN Genomes (TILLING), a reverse genetic approach, can take advantage of conventional mutation induction and high-through mutation methods, providing the capability of recovering mutations from any genetic regions and discover novel phenotypes [84]. For instance, TILLING was used to induce homozygous mutations in two waxy genes (granule-bound starch synthase I, or GBSSI) in a progenitor of wheat that has the null waxy genotype [84]. Forward and reverse genetic screening can also be combined. After conducting reverse genetic screening in *Lotus japon* using TILLING, forward genetic screening was used to recover the alleles from a set of 275 cultivars with a 10-fold reduction of the cost compared with whole-genome screening methods [85].

4.4. Genomic Selection

Trait discovery using QTL analysis, GWAS or reverse genetics is not essential for genomics-based breeding. Particularly when targeting polygenic agronomic traits such as yield, minor effect alleles can be difficult to detect and assess using these methods. When confronted with complex traits that are difficult to introgress systematically, GS present an additional breeding approach. GS relies on calculating the genomic estimated breeding values (GEBV) for sets of variants [86] based on a genotyped and phenotyped training population. By combining the entire SNP marker sets in the predicted model and preventing biased marker effects from variations in small-effect QTL, GS overcomes inefficient translation of QTL analysis results from biparental mapping populations to breeding and insufficient capability of statistical approaches to identify polygenic loci [87,88]. The combination of GS with automated phenotyping techniques can further promote prediction accuracy of GEBV, shortening the breeding cycle [88]. Based on computational simulations, GS based genomics-assisted breeding allows a four-year reduction in the breeding cycle of pasture grass *Lolium perenne* compared with traditional breeding [89]. Additionally, GS in cassava showed a significant increase in predicted genetic gains from 39.42% to 73.96%, comparing with the phenotypic selection [90]. Genotyping-by-sequencing (GBS) allows low-cost *de novo* genotyping

of crop plants [91,92] and can help to develop accurate GS models in crops with large genomes that are costly to genotype with whole genome sequencing at the population level [93]. GBS was used in elite wheat breeding lines to develop precise GS models, estimating genotypes with high yield and stem rust resistance [93,94]. GBS also delivered 55,000 SNP markers with high prediction accuracy from various maize breeding lines that were used for GS model construction [95].

4.5. Beyond the Gene: Targeting Cis-Regulatory Elements for Crop Breeding

Cis-regulatory elements (CREs) such as promotors and enhancers regulate gene expression and may contain close to half of all variants influencing traits [96]. In crops, domestication traits are often caused by variants in CREs [97]. For instance, Wang et al. showed that a mutation in a rice CRE achieved the production of slender rice without decreasing yield production, by reducing the repression of the gene *GRAIN WIDTH 7* [98]. Targeting CREs for breeding can be advantageous when the aim is not to knock out a gene entirely but to reduce or increase expression. Although CREs are not expressed like genes, making them more difficult to study, they are associated with open chromatin, which facilitates protein-binding. Open chromatin can be identified through DNase I hypersensitivity mapping [99], ATAC-seq [100] and ChIP-seq [101] experiments, which helps predict candidate CREs. Recently developed laboratory approaches combine chromatin signature detection and genome editing to allow prediction, validation and genome-wide functional assessment of CREs. For example, by editing ChIP-seq identified candidate sites, 73 enhancers were detected in humans, using an approach readily transferable to plants [102]. Additionally, bioinformatics approaches such as word-counting across the different promoter sequences [103] and analysis of sequence conservation [104] are used for regulatory element detection. Due to these methods, identifying *cis*-regulatory elements has become easier, leading to a growing body of knowledge on mammalian and, to a lesser extent, plant *cis*-regulatory elements.

As our high-throughput CRE detection ability grows, the challenge for breeders is knowing which CRE to target. This is difficult because knowledge of the functional impact of CREs in plants is scant [98]. Integrated databases such as the Plant *Cis*-Acting Regulatory Elements (Plant CARE) database provide comprehensive resources to study plant CRE function [105]. However, the roles of specific CREs in regulatory networks are largely unknown and whether editing the sequence of a CRE will affect the expression of the target gene can only be accurately determined experimentally. The experiments necessary to characterise CREs in this way have been conducted on genes in rice [106,107], generating a combined mutant library of almost 100,000 independent lines. By producing a CRE mutant library with a similar approach and obtaining expression data from the mutant lines, CREs associated with agronomic traits could be identified. When trait-associated CREs are known, an allelic series produced with genome editing can rapidly create stepwise variation in a target trait. For example, using Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/Cas9 genome editing, Rodríguez-Leal et al. obtained stepwise variation in tomato seed compartment (locule) numbers, a yield-related trait, by targeting the promoters of the inflorescence architecture gene *WUSCHEL* (*WUS*) and *CLAVATA* (*CLV3*) [108].

5. Applying Machine Learning to Crop Breeding

Machine learning (ML) allows algorithms to interpret data by learning patterns through experience. For large, diverse and formless datasets, such as those generated by photo imaging or sequencing, ML can provide substantial advantages over other analytical approaches [109]. With the aid of ML, crop breeders can efficiently phenotype plants and mine diverse datasets for patterns such as associations between DNA sequences and traits.

5.1. High Throughput Crop Phenotyping

Plant phenotyping is the measurement of functional or structural traits from the cellular level to the organism level and is essential for association studies and crop improvement [110]. With the

intensive development of genomics research and sequencing techniques, there is an increasing demand for plant phenotypes to help understand genomic data. Conventional phenotyping is often a bottleneck because it is subjective, error-prone, labour-intensive and time-consuming, limiting the number of traits, plants and environments that can be sampled [111]. Advances in measuring technologies (high-throughput imaging and automatic sensors) and ML allow the establishment of robotic high-throughput phenotyping, overcoming the shortcomings of traditional human-based phenotyping by allowing rapid generation of phenotypical features and features across large populations. High-throughput phenotyping has four main elements, namely detection by imaging or sensors, phenotypical data classification, feature quantification and prediction based on specific models or algorithms [112]. Phenotyping using ML has been applied in stress phenotyping and monitoring of diseases. A real-time ML-based high-throughput phenotyping method was developed to assess the severity of iron deficiency chlorosis from a total of 4366 soybeans from representative canopies [113]. Using linear discriminant analysis (LDA) and multi-class support vector machines (SVM) ML algorithms, the collected phenotypic data were used to train the best classification model to predict iron deficiency chlorosis stress severity in soybean, which is useful to measure the real-time severity in the soybean field. Another recent study applied a deep ML-based phenotyping method using an unsupervised identification explanation mechanism to measure, determine and classify various stress severity of foliar stresses in *Glycine max* including bacterial and fungal diseases and nutrient deficiency [114]. Although ML has been successfully applied in crop genomics and crop phenotyping, several challenges remain. ML modelling requires large datasets for training and model construction. A small training set can lead to statistically insignificant and problematic prediction [115] but a large dataset can be costly and time-consuming to acquire, as measurements on crops can often only be taken once per growth cycle. High-throughput ML based phenotyping therefore remains limited to certain research institutes and commercial companies [116]. Further reduction of purchasing and operating costs is required to make ML-based phenotyping widely usable on the farm of the future.

5.2. Machine Learning in Crop Genomics Research

ML plays a role in many areas of genomics research, including genome assembly, the iterative inference of gene regulatory networks and identifying true SNPs in polyploid plants. A representative list of ML algorithms and related open-source R packages relevant for data analysis in plants is provided by Ma et al. [117]. ML can be used to improve assemblies of polyploid genomes with complex genome redundancies [117]. A complete genome assembly and annotation is the foundation to track the genetic variations within a plant species and understand plant gene function and structure, which are crucial within the crop trait discovery pipeline (Figure 1). Assembling highly redundant genomes is a challenge for a non-ML-based assembly approaches that use a linear algorithm to assemble repetitive sequence regions. To overcome this limitation, an ML method was used to detect assembly errors and generate a high-quality assembly of bread wheat (*Triticum aestivum*) [118]. ML is also deployed by the RNA-seq mapping tool Portcullis to differentiate between real and artificial splicing junctions, which has helped annotate the bread wheat genome [119].

Inferring the relationships between regulatory elements and genes is a promising field for identifying previously unknown candidates for crop improvement. Based only on gene co-expression levels, a regulatory network constructed in silico is limited because the association between genes may not accurately reflect shared gene regulation [120]. Consequently, an ML-based method that can incorporate various kinds of regulatory signals from different data sources has become popular for interactive inference of gene regulatory networks [117]. For instance, by analysing data on transcription factor binding, conserved sequences, gene expression and chromatin modification data, the transcriptional regulatory network in *Drosophila melanogaster* involving 300,000 regulatory edges and over 12,000 targeted genes could be predicted [121].

SNPs are the most frequent class of variations in plant genomes [122,123]. However, challenges remain for SNP discovery in polyploid plants [124]. Korani et al. [125] developed an ML-based

analysis tool called SNP-ML using neural networks and tree bagging models to efficiently filter false positive SNPs. They demonstrated that SNP-ML could be used to detect SNP variation and select true SNPs with over 98% accuracy in simulated SNP variant data of peanut, cotton and strawberry. Neural networks have also been deployed for SNP calling in error-prone long reads, which is particularly challenging in diploid or polyploid samples as errors must be distinguished from genuine heterozygous variants [126].

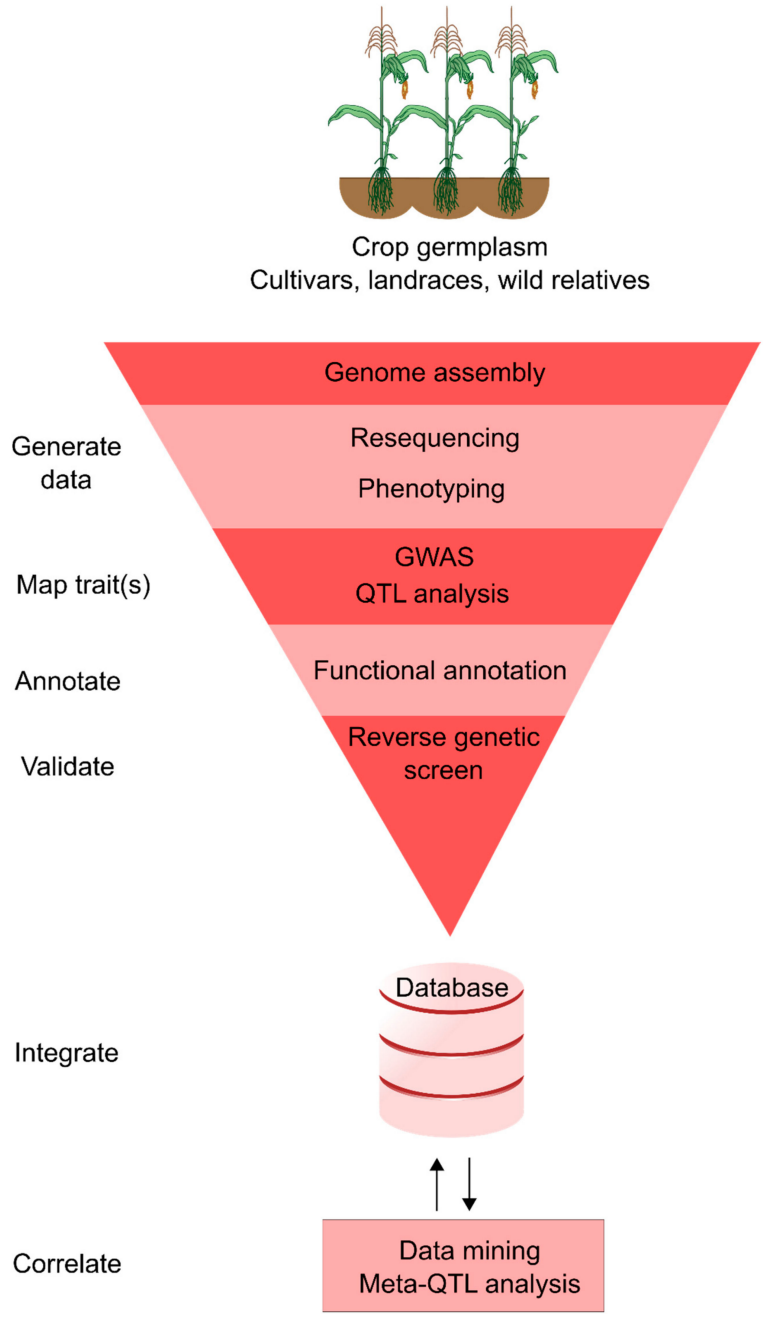

Figure 1. Schematic overview of a crop trait discovery pipeline.

6. Genome Editing of Crops and Bioinformatics Challenges in Guide RNA Design

Breeding relies on generating novel combinations of alleles and this can be achieved by harnessing natural variation from germplasm collections or by generating novel variation. Advanced breeding methods to generate DNA mutations such as irradiation or chemical mutagens are commonly

used [127]. However, these methods are hampered by the high rate of background mutations, some of which can be deleterious and need to be removed with multiple time-consuming rounds of breeding. In contrast to conventional breeding methods (Figure 2), CRISPR/Cas does not require substantial crossing to fix traits and remove deleterious background mutations. The CRISPR/Cas system relies on a guide RNA (gRNA) to target the Cas protein to DNA sites for cleavage by locating a matching ~20 bp sequence and a protospacer adjacent motif (PAM), which is specific to the Cas protein used [128]. The Cas protein induces a double-strand break at the target site, allowing gene knock-out via mutations arising during nonhomologous end joining and gene knock-in via a donor DNA template and homology-directed repair. In the last five years, the CRISPR/Cas system has been efficiently applied in a variety of essential food crops (reviewed in [129]) and is expected to have a major impact on agriculture [130,131]. The remaining challenges for genome editors are improving protoplast transformation [132], increasing the efficiency of gene targeting using homology-directed repair [133] and optimisingthe bioinformatics tools for gRNA design with minimal off-target effects.

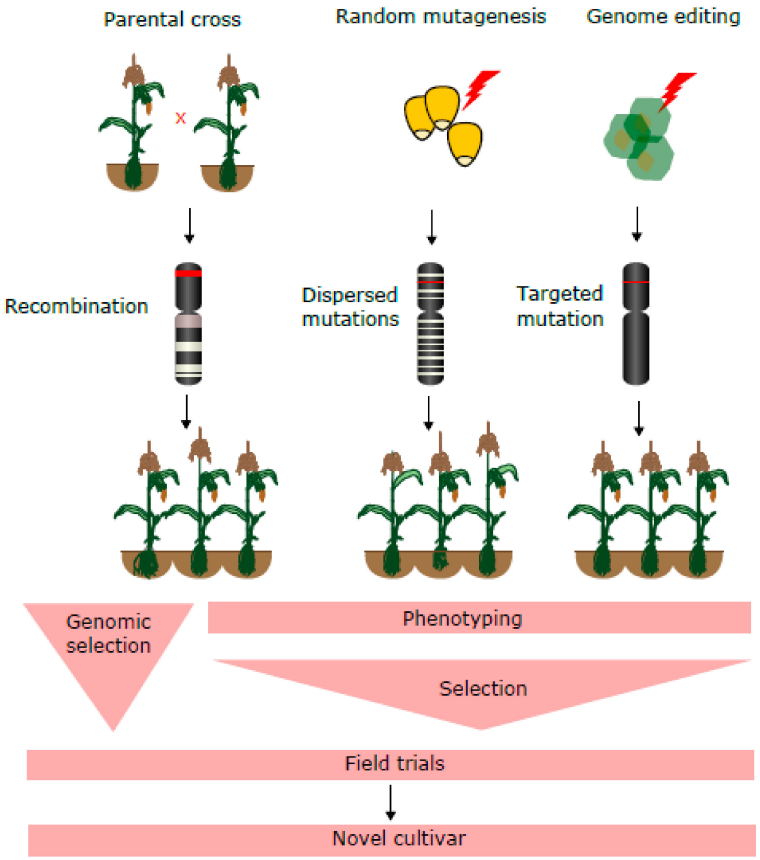

Figure 2. Generating a novel crop cultivar using a parental cross, random mutagenesis and genome editing. Changed genomic regions on a representative chromosome are shown, with regions harbouring beneficial mutations in red. Plant populations shown to harbour more phenotypic changes when using a parental cross or random mutagenesis and are more uniform when using genome editing, likely varying only in the target trait(s). When using genomic selection to complement traditional breeding, the phenotyping process can be replaced with a faster genotyping step and plants selected based on breeding values calculated from a phenotyped and genotyped training population.

Bioinformatics tools are essential for the optimal design of gRNAs that facilitate efficient and specific CRISPR/Cas gene editing. The two crucial requirements for gRNA design are that the gRNA has both high binding affinity to the target site and is specific, with few off-target effects. Using human cell lines, it was shown that guanines are at the −1 and −2 PAM-proximal positions increase binding

efficiency, whereas thymines at the +4/−4 PAM-proximal positions decrease efficiency [134]. An assay on human cells investigated the effect of nucleotide sequence and epigenetic parameters on gRNA efficiency, finding that both locus accessibility and the sequence composition affect efficiency [135]. Similar studies of gRNA binding efficiency are lacking in plants. However, early assessments indicate that the base preferences identified in human cells are not shared by plants [136]. The specificity of gRNA designs is difficult to predict based on sequence similarity alone and the results of off-target prediction are inconsistent between various tools [137]. Although there are over 50 tools for gRNA design publicly available (http://omictools.com/crispr-cas9-category) [138], only CRISPR-P [139] and CRISPR-Plant [140] are designed for plants. CRISPR-P is the more sophisticated tool, supporting design for 49 plant species and providing secondary structure analysis and microhomology scores for guide designs. However, the gRNA activity evaluations and specific searches for knock-in or knock-out designs available for non-plant model systems in tools such as the CRISPR-ERA web server [141] as well as learning-based design tools such as sgRNA Designer [142] are not yet available for plants. Because the interactions between Cas proteins, gRNA, DNA and chromatin are likely to differ to some degree in plants [136], it will be important to incorporate this information into plant gRNA design tools. In addition, plant genomes are highly redundant which may make it difficult to generate unique gRNAs for single target sites. Moreover, for minor crops with few available genomic resources, or highly genetically diverse germplasm, it can be difficult to predict gRNA activity because of SNPs between the reference genome and the target individual [143]. As additional endonucleases with different activity from Cas9 are added to the CRISPR/Cas toolkit and more empirical data on endonuclease activity in plants becomes available, bioinformatics tools will be able to tailor gRNAs to target more genomic sites at higher accuracy in crops.

7. Conclusions

Agriculture faces substantial challenges in harnessing the deluge of genomic data of diverse origins and formats for crop improvement. To overcome these challenges, novel breeding methods and bioinformatics tools must be used to translate genomic data into gains in crop yield and yield stability. To accelerate the detection of robust gene-trait associations, researchers can apply meta-QTL analyses, GWAS and genetic screens. While genome editing offers a valuable approach to rapidly introduce beneficial mutations into elite cultivars, GS increases selection efficiency without requiring knowledge of underlying genetic drivers. ML algorithms can take advantage of high-throughput phenotyping and genomic data to further automate parts of the gene discovery pipeline such as genome annotation and image interpretation that remain particularly challenging. By applying novel technologies and methods in concert, future plant breeding can achieve the crop improvement rate required to ensure food security.

Author Contributions: All authors co-wrote the manuscript.

Acknowledgments: H.H. thanks the China Scholarship Council for supporting his studies at the University of Western Australia. A.S. was supported by an IPRS awarded by the Australian government. This work is funded by the Australian Research Council (Projects LP160100030, LP140100537 and LP130100925).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Smit, E.; Nieto, F.J.; Crespo, C.J.; Mitchell, P. Estimates of animal and plant protein intake in US adults: Results from the Third National Health and Nutrition Examination Survey, 1988–1991. *J. Acad. Nutr. Diet.* **1999**, *99*, 813–820. [CrossRef]
2. Ulijaszek, S.J. Human dietary change. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **1991**, *334*, 271–279. [CrossRef] [PubMed]
3. Tilman, D.; Balzer, C.; Hill, J.; Befort, B.L. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 20260–20264. [CrossRef] [PubMed]

4. Godfray, H.C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food security: The challenge of feeding 9 billion people. *Science* **2010**, *327*, 812–818. [CrossRef] [PubMed]
5. Perez-de-Castro, A.M.; Vilanova, S.; Cañizares, J.; Pascual, L.; M Blanca, J.; J Diez, M.; Prohens, J.; Picó, B. Application of genomic tools in plant breeding. *Curr. Genom.* **2012**, *13*, 179–195. [CrossRef] [PubMed]
6. Abberton, M.; Batley, J.; Bentley, A.; Bryant, J.; Cai, H.; Cockram, J.; de Oliveira, A.C.; Cseke, L.J.; Dempewolf, H.; De Pace, C.; et al. Global agricultural intensification during climate change: A role for genomics. *Plant Biotechnol. J.* **2016**, *14*, 1095–1098. [CrossRef] [PubMed]
7. Edwards, D.; Batley, J.; Snowdon, R.J. Accessing complex crop genomes with next-generation sequencing. *Theor. Appl. Genet.* **2013**, *126*, 1–11. [CrossRef] [PubMed]
8. IRGSP. The map-based sequence of the rice genome. *Nature* **2005**, *436*, 793–800. [CrossRef]
9. Velasco, R.; Zharkikh, A.; Troggio, M.; Cartwright, D.A.; Cestaro, A.; Pruss, D.; Pindo, M.; Fitzgerald, L.M.; Vezzulli, S.; Reid, J.; et al. A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS ONE* **2007**, *2*, e1326. [CrossRef] [PubMed]
10. Huang, S.; Li, R.; Zhang, Z.; Li, L.; Gu, X.; Fan, W.; Lucas, W.J.; Wang, X.; Xie, B.; Ni, P.; et al. The genome of the cucumber, *Cucumis sativus* L. *Nat. Genet.* **2009**, *41*, 1275–1281. [CrossRef] [PubMed]
11. Goodwin, S.; McPherson, J.D.; McCombie, W.R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **2016**, *17*, 333–351. [CrossRef] [PubMed]
12. Michael, T.P.; Jackson, S. The First 50 Plant Genomes. *Plant Genome* **2013**, *6*, 1–7. [CrossRef]
13. Mousavi-Derazmahalleh, M.; Bayer, P.E.; Hane, J.K.; Babu, V.; Nguyen, H.T.; Nelson, M.N.; Erskine, W.; Varshney, R.K.; Papa, R.; Edwards, D. Adapting legume crops to climate change using genomic approaches. *Plant Cell Environ* **2018**. [CrossRef] [PubMed]
14. Dwivedi, S.L.; Scheben, A.; Edwards, D.; Spillane, C.; Ortiz, R. Assessing and exploiting functional diversity in germplasm pools to enhance abiotic stress adaptation and yield in cereals and food legumes. *Front. Plant Sci.* **2017**, *8*, 1461. [CrossRef] [PubMed]
15. Moore, J.H.; Asselbergs, F.W.; Williams, S.M. Bioinformatics challenges for genome-wide association studies. *Bioinformatics* **2010**, *26*, 445–455. [CrossRef] [PubMed]
16. Batley, J.; Edwards, D. The application of genomics and bioinformatics to accelerate crop improvement in a changing climate. *Curr. Opin. Plant Biol.* **2016**, *30*, 78–81. [CrossRef] [PubMed]
17. Lawrence, M.; Huber, W.; Pages, H.; Aboyoun, P.; Carlson, M.; Gentleman, R.; Morgan, M.T.; Carey, V.J. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **2013**, *9*, e1003118. [CrossRef] [PubMed]
18. Pabinger, S.; Dander, A.; Fischer, M.; Snajder, R.; Sperk, M.; Efremova, M.; Krabichler, B.; Speicher, M.R.; Zschocke, J.; Trajanoski, Z. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief. Bioinform.* **2014**, *15*, 256–278. [CrossRef] [PubMed]
19. 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]
20. Sedlazeck, F.J.; Lee, H.; Darby, C.A.; Schatz, M.C. Piercing the dark matter: Bioinformatics of long-range sequencing and mapping. *Nat. Rev. Genet.* **2018**, *6*, 329–346. [CrossRef] [PubMed]
21. Ong, Q.; Nguyen, P.; Phuong Thao, N.; Le, L. Bioinformatics approach in plant genomic research. *Curr. Genom.* **2016**, *17*, 368–378. [CrossRef] [PubMed]
22. Grierson, C.S.; Barnes, S.R.; Chase, M.W.; Clarke, M.; Grierson, D.; Edwards, K.J.; Jellis, G.J.; Jones, J.D.; Knapp, S.; Oldroyd, G.; et al. One hundred important questions facing plant science research. *New Phytol.* **2011**, *192*, 6–12. [CrossRef] [PubMed]
23. Matthews, D.E.; Carollo, V.L.; Lazo, G.R.; Anderson, O.D. GrainGenes, the genome database for small-grain crops. *Nucleic Acids Res.* **2003**, *31*, 183–186. [CrossRef] [PubMed]
24. Tello-Ruiz, M.K.; Naithani, S.; Stein, J.C.; Gupta, P.; Campbell, M.; Olson, A.; Wei, S.R.; Preece, J.; Geniza, M.J.; Jiao, Y.P.; et al. Gramene 2018: Unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res.* **2018**, *46*, D1181–D1189. [CrossRef] [PubMed]
25. Scheben, A.; Batley, J.; Edwards, D. Revolution in genotyping platforms for crop improvement. In *Advances in Biochemical Engineering/Biotechnology*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 1–16.
26. Moose, S.P.; Mumm, R.H. Molecular plant breeding as the foundation for 21st century crop improvement. *Plant Physiol.* **2008**, *147*, 969–977. [CrossRef] [PubMed]

27. Watson, A.; Ghosh, S.; Williams, M.J.; Cuddy, W.S.; Simmonds, J.; Rey, M.D.; Hatta, M.A.M.; Hinchliffe, A.; Steed, A.; Reynolds, D.; et al. Speed breeding is a powerful tool to accelerate crop research and breeding. *Nat. Plants* **2018**, *4*, 23–29. [CrossRef] [PubMed]
28. Santos, R.; Algar, A.; Field, R.; Mayes, S. Integrating GIScience and Crop Science datasets: A study involving genetic, geographic and environmental data. *PeerJ Preprints* **2017**, *5*, e2248v2244. [CrossRef]
29. Evans, K.; Jung, S.; Lee, T.; Brutcher, L.; Cho, I.; Peace, C.; Main, D. Addition of a breeding database in the Genome Database for Rosaceae. *Database* **2013**, *2013*, bat078. [CrossRef] [PubMed]
30. Li, C.; Lin, F.; An, D.; Wang, W.; Huang, R. Genome Sequencing and Assembly by Long Reads in Plants. *Genes* **2017**, *9*, 6. [CrossRef] [PubMed]
31. Vlk, D.; Repkova, J. Application of next-generation sequencing in plant breeding. *Czech J. Genet. Plant* **2017**, *53*, 89–96. [CrossRef]
32. Chen, K.; Ji, F.; Yuan, S.J.; Hao, W.T.; Wang, W.; Hu, Z.H. The performance of activated sludge exposed to arsanilic acid and amprolium hydrochloride in sequencing batch reactors. *Int. Biodeterior. Biodegrad.* **2017**, *116*, 260–265. [CrossRef]
33. Jiao, W.-B.; Schneeberger, K. The impact of third generation genomic technologies on plant genome assembly. *Curr. Opin. Plant Biol.* **2017**, *36*, 64–70. [CrossRef] [PubMed]
34. VanBuren, R.; Bryant, D.; Edger, P.P.; Tang, H.; Burgess, D.; Challabathula, D.; Spittle, K.; Hall, R.; Gu, J.; Lyons, E.; et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. *Nature* **2015**, *527*, 508–511. [CrossRef] [PubMed]
35. Stankova, H.; Hastie, A.R.; Chan, S.; Vrana, J.; Tulpova, Z.; Kubalakova, M.; Visendi, P.; Hayashi, S.; Luo, M.; Batley, J.; et al. BioNano genome mapping of individual chromosomes supports physical mapping and sequence assembly in complex plant genomes. *Plant Biotechnol. J.* **2016**, *14*, 1523–1531. [CrossRef] [PubMed]
36. Van Berkum, N.L.; Lieberman-Aiden, E.; Williams, L.; Imakaev, M.; Gnirke, A.; Mirny, L.A.; Dekker, J.; Lander, E.S. Hi-C: A method to study the three-dimensional architecture of genomes. *J. Vis. Exp.* **2010**, *39*, e1869. [CrossRef] [PubMed]
37. Mascher, M.; Gundlach, H.; Himmelbach, A.; Beier, S.; Twardziok, S.O.; Wicker, T.; Radchuk, V.; Dockter, C.; Hedley, P.E.; Russell, J.; et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature* **2017**, *544*, 427–433. [CrossRef] [PubMed]
38. Michael, T.P.; Jupe, F.; Bemm, F.; Motley, S.T.; Sandoval, J.P.; Lanz, C.; Loudet, O.; Weigel, D.; Ecker, J.R. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat. Commun.* **2018**, *9*, 541. [CrossRef] [PubMed]
39. Goodwin, S.; Gurtowski, J.; Ethe-Sayers, S.; Deshpande, P.; Schatz, M.C.; McCombie, W.R. Oxford Nanopore sequencing, hybrid error correction and de novo assembly of a eukaryotic genome. *Genome Res.* **2015**, *25*, 1750–1756. [CrossRef] [PubMed]
40. Med.stanford.edu. Stanford Medicine Sequencing Service Rates. Available online: http://med.stanford.edu/gssc/rates.html (accessed on 27 May 2018).
41. Cgrb.oregonstate.edu. Illumina HiSeq 3000 Service Fees. Available online: http://cgrb.oregonstate.edu/core/illumina-hiseq-3000/illumina-hiseq-3000-service-fees (accessed on 27 May 2018).
42. Allseq.com. General overview of Illumina Sequencing. Available online: http://allseq.com/knowledge-bank/sequencing-platforms/illumina/ (accessed on 27 May 2018).
43. Gordon, D.; Huddleston, J.; Chaisson, M.J.P.; Hill, C.M.; Kronenberg, Z.N.; Munson, K.M.; Malig, M.; Raja, A.; Fiddes, I.; Hillier, L.W.; et al. Long-read sequence assembly of the gorilla genome. *Science* **2016**, *352*, aae0344. [CrossRef] [PubMed]
44. Washington.edu. University of Washington PacBio Sequencing Services. Available online: https://pacbio.gs.washington.edu/ (accessed on 27 May 2018).
45. Laver, T.; Harrison, J.; O'Neill, P.A.; Moore, K.; Farbos, A.; Paszkiewicz, K.; Studholme, D.J. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomol. Detect. Quantif.* **2015**, *3*, 1–8. [CrossRef] [PubMed]
46. Jain, M.; Koren, S.; Miga, K.H.; Quick, J.; Rand, A.C.; Sasani, T.A.; Tyson, J.R.; Beggs, A.D.; Dilthey, A.T.; Fiddes, I.T.; et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **2018**, *36*, 338–345. [CrossRef] [PubMed]

47. George, S.; Pankhurst, L.; Hubbard, A.; Votintseva, A.; Stoesser, N.; Sheppard, A.E.; Mathers, A.; Norris, R.; Navickaite, I.; Eaton, C.; et al. Resolving plasmid structures in Enterobacteriaceae using the MinION nanopore sequencer: Assessment of MinION and MinION/Illumina hybrid data assembly approaches. *Microb. Genom.* **2017**, *3*, e000118. [CrossRef] [PubMed]
48. Nanoporetech.com. Available online: https://nanoporetech.com (accessed on 27 May 2018).
49. Schmidt, M.H.W.; Vogel, A.; Denton, A.K.; Istace, B.; Wormit, A.; van de Geest, H.; Bolger, M.E.; Alseekh, S.; Mass, J.; Pfaff, C.; et al. *De novo* assembly of a new *Solanum pennellii* accession using nanopore sequencing. *Plant Cell* **2017**, *29*, 2336–2348. [CrossRef] [PubMed]
50. Pareek, C.S.; Smoczynski, R.; Tretyn, A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* **2011**, *52*, 413–435. [CrossRef] [PubMed]
51. Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J.; Wheeler, D.L. GenBank. *Nucleic Acids Res.* **2008**, *36*, D25. [CrossRef] [PubMed]
52. Kanz, C.; Aldebert, P.; Althorpe, N.; Baker, W.; Baldwin, A.; Bates, K.; Browne, P.; van den Broek, A.; Castro, M.; Cochrane, G.; et al. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **2005**, *33*, D29–D33. [CrossRef] [PubMed]
53. Duvick, J.; Fu, A.; Muppirala, U.; Sabharwal, M.; Wilkerson, M.D.; Lawrence, C.J.; Lushbough, C.; Brendel, V. PlantGDB: A resource for comparative plant genomics. *Nucleic Acids Res.* **2007**, *36*, D959–D965. [CrossRef] [PubMed]
54. Goodstein, D.M.; Shu, S.; Howson, R.; Neupane, R.; Hayes, R.D.; Fazo, J.; Mitros, T.; Dirks, W.; Hellsten, U.; Putnam, N.; et al. Phytozome: A comparative platform for green plant genomics. *Nucleic Acids Res.* **2012**, *40*, D1178–D1186. [CrossRef] [PubMed]
55. Lai, K.; Lorenc, M.T.; Edwards, D. Genomic databases for crop improvement. *Agronomy* **2012**, *2*, 62. [CrossRef]
56. Hassani-Pak, K.; Rawlings, C. Knowledge Discovery in Biological Databases for Revealing Candidate Genes Linked to Complex Phenotypes. *J. Integr. Bioinform.* **2017**, *14*. [CrossRef] [PubMed]
57. Hassani-Pak, K.; Castellote, M.; Esch, M.; Hindle, M.; Lysenko, A.; Taubert, J.; Rawlings, C. Developing integrated crop knowledge networks to advance candidate gene discovery. *Appl. Transl. Genom.* **2016**, *11*, 18–26. [CrossRef] [PubMed]
58. Yuan, Y.; Scheben, A.; Chan, C.K.; Edwards, D. Databases for wheat genomics and crop improvement. In *Methods in Molecular Biology*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 1679, pp. 277–291.
59. Furbank, R.T.; Tester, M. Phenomics–technologies to relieve the phenotyping bottleneck. *Trends Plant Sci.* **2011**, *16*, 635–644. [CrossRef] [PubMed]
60. Mackay, T.F.; Stone, E.A.; Ayroles, J.F. The genetics of quantitative traits: Challenges and prospects. *Nat. Rev. Genet.* **2009**, *10*, 565–577. [CrossRef] [PubMed]
61. Mace, E.; Jordan, D. Integrating sorghum whole genome sequence information with a compendium of sorghum QTL studies reveals uneven distribution of QTL and of gene-rich regions with significant implications for crop improvement. *Theor. Appl. Genet.* **2011**, *123*, 169–191. [CrossRef] [PubMed]
62. Veyrieras, J.B.; Goffinet, B.; Charcosset, A. MetaQTL: A package of new computational methods for the meta-analysis of QTL mapping experiments. *BMC Bioinform.* **2007**, *8*, 49. [CrossRef] [PubMed]
63. Kumasaka, N.; Knights, A.J.; Gaffney, D.J. Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nat. Genet.* **2016**, *48*, 206. [CrossRef] [PubMed]
64. Tecle, I.Y.; Menda, N.; Buels, R.M.; van der Knaap, E.; Mueller, L.A. solQTL: A tool for QTL analysis, visualization and linking to genomes at SGN database. *BMC Bioinform.* **2010**, *11*, 525. [CrossRef] [PubMed]
65. Liu, R.X.; Zhang, H.; Zhao, P.; Zhang, Z.X.; Liang, W.K.; Tian, Z.G.; Zheng, Y.L. Mining of candidate maize genes for nitrogen use efficiency by integrating gene expression and QTL data. *Plant Mol. Biol. Report.* **2012**, *30*, 297–308. [CrossRef]
66. Said, J.I.; Lin, Z.; Zhang, X.; Song, M.; Zhang, J. A comprehensive meta QTL analysis for fiber quality, yield, yield related and morphological traits, drought tolerance and disease resistance in tetraploid cotton. *BMC Genom.* **2013**, *14*, 776. [CrossRef] [PubMed]
67. Qi, Z.-M.; Wu, Q.; Han, X.; Sun, Y.-N.; Du, X.-Y.; Liu, C.-Y.; Jiang, H.-W.; Hu, G.-H.; Chen, Q.-S. Soybean oil content QTL mapping and integrating with meta-analysis method for mining genes. *Euphytica* **2011**, *179*, 499–514. [CrossRef]

68. Hanocq, E.; Laperche, A.; Jaminon, O.; Laine, A.-L.; Le Gouis, J. Most significant genome regions involved in the control of earliness traits in bread wheat, as revealed by QTL meta-analysis. *Theor. Appl. Genet.* **2007**, *114*, 569–584. [CrossRef] [PubMed]
69. Borevitz, J.O.; Nordborg, M. The impact of genomics on the study of natural variation in Arabidopsis. *Plant Physiol.* **2003**, *132*, 718–725. [CrossRef] [PubMed]
70. Bergelson, J.; Roux, F. Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. *Nat. Rev. Genet.* **2010**, *11*, 867–879. [CrossRef] [PubMed]
71. Huang, X.; Han, B. Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* **2014**, *65*, 531–551. [CrossRef] [PubMed]
72. Korte, A.; Farlow, A. The advantages and limitations of trait analysis with GWAS: A review. *Plant Methods* **2013**, *9*, 29. [CrossRef] [PubMed]
73. Huang, X.; Wei, X.; Sang, T.; Zhao, Q.; Feng, Q.; Zhao, Y.; Li, C.; Zhu, C.; Lu, T.; Zhang, Z.; et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **2010**, *42*, 961–967. [CrossRef] [PubMed]
74. Sonah, H.; O'Donoughue, L.; Cober, E.; Rajcan, I.; Belzile, F. Identification of loci governing eight agronomic traits using a GBS-GWAS approach and validation by QTL mapping in soya bean. *Plant Biotechnol. J.* **2015**, *13*, 211–221. [CrossRef] [PubMed]
75. Tian, F.; Bradbury, P.J.; Brown, P.J.; Hung, H.; Sun, Q.; Flint-Garcia, S.; Rocheford, T.R.; McMullen, M.D.; Holland, J.B.; Buckler, E.S. Genome-wide association study of leaf architecture in the maize nested association mapping population. *Nat. Genet.* **2011**, *43*, 159–162. [CrossRef] [PubMed]
76. Gacek, K.; Bayer, P.E.; Bartkowiak-Broda, I.; Szala, L.; Bocianowski, J.; Edwards, D.; Batley, J. Genome-wide association study of genetic control of seed fatty acid biosynthesis in *Brassica napus*. *Front. Plant Sci.* **2017**, *7*, 2062. [CrossRef] [PubMed]
77. Purcell, S.; Neale, B.; Todd-Brown, K.; Thomas, L.; Ferreira, M.A.; Bender, D.; Maller, J.; Sklar, P.; de Bakker, P.I.; Daly, M.J.; et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **2007**, *81*, 559–575. [CrossRef] [PubMed]
78. Ma, C.; Blackwell, T.; Boehnke, M.; Scott, L.J.; GoT2D Investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet. Epidemiol.* **2013**, *37*, 539–550. [CrossRef] [PubMed]
79. Bradbury, P.J.; Zhang, Z.; Kroon, D.E.; Casstevens, T.M.; Ramdoss, Y.; Buckler, E.S. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **2007**, *23*, 2633–2635. [CrossRef] [PubMed]
80. Lipka, A.E.; Tian, F.; Wang, Q.; Peiffer, J.; Li, M.; Bradbury, P.J.; Gore, M.A.; Buckler, E.S.; Zhang, Z. GAPIT: Genome association and prediction integrated tool. *Bioinformatics* **2012**, *28*, 2397–2399. [CrossRef] [PubMed]
81. Jankowicz-Cieslak, J.; Till, B.J. Forward and reverse genetics in crop breeding. In *Advances in Plant Breeding Strategies: Breeding, Biotechnology and Molecular Tools*; Al-Khayri, J.M., Jain, S.M., Johnson, D.V., Eds.; Springer: Heidelberg, Germany, 2015; Volume 1, pp. 215–240.
82. Sessions, A.; Burke, E.; Presting, G.; Aux, G.; McElver, J.; Patton, D.; Dietrich, B.; Ho, P.; Bacwaden, J.; Ko, C.; et al. A high-throughput Arabidopsis reverse genetics system. *Plant Cell* **2002**, *14*, 2985–2994. [CrossRef] [PubMed]
83. Henry, I.M.; Nagalakshmi, U.; Lieberman, M.C.; Ngo, K.J.; Krasileva, K.V.; Vasquez-Gross, H.; Akhunova, A.; Akhunov, E.; Dubcovsky, J.; Tai, T.H.; et al. Efficient genome-wide detection and cataloging of EMS-induced mutations using exome capture and next-generation sequencing. *Plant Cell* **2014**, *26*, 1382–1397. [CrossRef] [PubMed]
84. Slade, A.J.; Fuerstenberg, S.I.; Loeffler, D.; Steine, M.N.; Facciotti, D. A reverse genetic, nontransgenic approach to wheat crop improvement by TILLING. *Nat. Biotechnol.* **2005**, *23*, 75–81. [CrossRef] [PubMed]
85. Perry, J.A.; Wang, T.L.; Welham, T.J.; Gardner, S.; Pike, J.M.; Yoshida, S.; Parniske, M. A TILLING reverse genetics tool and a web-accessible collection of mutants of the legume *Lotus japonicus*. *Plant Physiol.* **2003**, *131*, 866–871. [CrossRef] [PubMed]
86. Jannink, J.L.; Lorenz, A.J.; Iwata, H. Genomic selection in plant breeding: From theory to practice. *Brief. Funct. Genom.* **2010**, *9*, 166–177. [CrossRef] [PubMed]

87. Bhat, J.A.; Ali, S.; Salgotra, R.K.; Mir, Z.A.; Dutta, S.; Jadon, V.; Tyagi, A.; Mushtaq, M.; Jain, N.; Singh, P.K.; et al. Genomic selection in the era of next generation sequencing for complex traits in plant breeding. *Front. Genet.* **2016**, *7*, 221. [CrossRef] [PubMed]
88. Heffner, E.L.; Sorrells, M.E.; Jannink, J.-L. Genomic selection for crop improvement. *Crop Sci.* **2009**, *49*, 1–12. [CrossRef]
89. Lin, Z.; Cogan, N.O.; Pembleton, L.W.; Spangenberg, G.C.; Forster, J.W.; Hayes, B.J.; Daetwyler, H.D. Genetic gain and inbreeding from genomic selection in a simulated commercial breeding program for perennial ryegrass. *Plant Genome* **2016**, *9*, 1–12. [CrossRef] [PubMed]
90. De Oliveira, E.J.; de Resende, M.D.V.; da Silva Santos, V.; Ferreira, C.F.; Oliveira, G.A.F.; da Silva, M.S.; de Oliveira, L.A.; Aguilar-Vildoso, C.I. Genome-wide selection in cassava. *Euphytica* **2012**, *187*, 263–276. [CrossRef]
91. Scheben, A.; Batley, J.; Edwards, D. Genotyping-by-sequencing approaches to characterize crop genomes: Choosing the right tool for the right application. *Plant Biotechnol. J.* **2017**, *15*, 149–161. [CrossRef] [PubMed]
92. Voss-Fels, K.; Snowdon, R.J. Understanding and utilizing crop genome diversity via high-resolution genotyping. *Plant Biotechnol. J.* **2016**, *14*, 1086–1094. [CrossRef] [PubMed]
93. Poland, J.; Endelman, J.; Dawson, J.; Rutkoski, J.; Wu, S.Y.; Manes, Y.; Dreisigacker, S.; Crossa, J.; Sanchez-Villeda, H.; Sorrells, M.; et al. Genomic Selection in Wheat Breeding using Genotyping-by-Sequencing. *Plant Genome* **2012**, *5*, 103–113. [CrossRef]
94. Rutkoski, J.E.; Poland, J.A.; Singh, R.P.; Huerta-Espino, J.; Bhavani, S.; Barbier, H.; Rouse, M.N.; Jannink, J.L.; Sorrells, M.E. Genomic Selection for Quantitative Adult Plant Stem Rust Resistance in Wheat. *Plant Genome* **2014**, *7*, 1–10. [CrossRef]
95. Crossa, J.; Beyene, Y.; Kassa, S.; Perez, P.; Hickey, J.M.; Chen, C.; de los Campos, G.; Burgueno, J.; Windhausen, V.S.; Buckler, E.; et al. Genomic prediction in maize breeding populations with genotyping-by-sequencing. *G3 Genes Genomes Genet.* **2013**, *3*, 1903–1926. [CrossRef] [PubMed]
96. Meyer, R.S.; Purugganan, M.D. Evolution of crop species: Genetics of domestication and diversification. *Nat. Rev. Genet.* **2013**, *14*, 840–852. [CrossRef] [PubMed]
97. Swinnen, G.; Goossens, A.; Pauwels, L. Lessons from domestication: Targeting *cis*-regulatory elements for crop improvement. *Trends Plant Sci.* **2016**, *21*, 506–515. [CrossRef] [PubMed]
98. Wang, S.; Li, S.; Liu, Q.; Wu, K.; Zhang, J.; Wang, S.; Wang, Y.; Chen, X.; Zhang, Y.; Gao, C.; et al. The *OsSPL16-GW7* regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.* **2015**, *47*, 949–954. [CrossRef] [PubMed]
99. Pauler, F.M.; Stricker, S.H.; Warczok, K.E.; Barlow, D.P. Long-range DNase I hypersensitivity mapping reveals the imprinted Igf2r and Air promoters share *cis*-regulatory elements. *Genome Res.* **2005**, *15*, 1379–1387. [CrossRef] [PubMed]
100. Buenrostro, J.D.; Wu, B.; Chang, H.Y.; Greenleaf, W.J. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Curr. Protoc. Mol. Biol.* **2015**, *109*, 21–29. [CrossRef] [PubMed]
101. Johnson, D.S.; Mortazavi, A.; Myers, R.M.; Wold, B. Genome-wide mapping of in vivo protein-DNA interactions. *Science* **2007**, *316*, 1497–1502. [CrossRef] [PubMed]
102. Korkmaz, G.; Lopes, R.; Ugalde, A.P.; Nevedomskaya, E.; Han, R.Q.; Myacheva, K.; Zwart, W.; Elkon, R.; Agami, R. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nat. Biotechnol.* **2016**, *34*, 192–198. [CrossRef] [PubMed]
103. Rombauts, S.; Florquin, K.; Lescot, M.; Marchal, K.; Rouzé, P.; Van de Peer, Y. Computational approaches to identify promoters and *cis*-regulatory elements in plant genomes. *Plant Physiol.* **2003**, *132*, 1162–1176. [CrossRef] [PubMed]
104. Van de Velde, J.; Heyndrickx, K.S.; Vandepoele, K. Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis. *Plant Cell* **2014**, *26*, 2729–2745. [CrossRef] [PubMed]
105. Lescot, M.; Déhais, P.; Thijs, G.; Marchal, K.; Moreau, Y.; Van de Peer, Y.; Rouzé, P.; Rombauts, S. PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucleic Acids Res.* **2002**, *30*, 325–327. [CrossRef] [PubMed]
106. Meng, X.B.; Yu, H.; Zhang, Y.F.; Zhuang, F.; Song, X.G.; Gao, S.S.; Gao, C.X.; Li, J.Y. Construction of a genome-wide mutant library in rice using CRISPR/Cas9. *Mol. Plant* **2017**, *10*, 1238–1241. [CrossRef] [PubMed]

107. Lu, Y.M.; Ye, X.; Guo, R.M.; Huang, J.; Wang, W.; Tang, J.Y.; Tan, L.T.; Zhu, J.K.; Chu, C.C.; Qian, Y.W. Genome-wide targeted mutagenesis in rice using the CRISPR/Cas9 system. *Mol. Plant* **2017**, *10*, 1242–1245. [CrossRef] [PubMed]
108. Rodríguez-Leal, D.; Lemmon, Z.H.; Man, J.; Bartlett, M.E.; Lippman, Z.B. Engineering quantitative trait variation for crop improvement by genome editing. *Cell* **2017**, *171*, 470–480. [CrossRef] [PubMed]
109. Libbrecht, M.W.; Noble, W.S. Machine learning applications in genetics and genomics. *Nat. Rev. Genet.* **2015**, *16*, 321–332. [CrossRef] [PubMed]
110. Walter, A.; Liebisch, F.; Hund, A. Plant phenotyping: From bean weighing to image analysis. *Plant Methods* **2015**, *11*, 14. [CrossRef] [PubMed]
111. Liu, N.; Koh, Z.X.; Goh, J.; Lin, Z.; Haaland, B.; Ting, B.P.; Ong, M.E.H. Prediction of adverse cardiac events in emergency department patients with chest pain using machine learning for variable selection. *BMC Med. Inform. Decis. Mak.* **2014**, *14*, 75. [CrossRef] [PubMed]
112. Singh, A.; Ganapathysubramanian, B.; Singh, A.K.; Sarkar, S. Machine learning for high-throughput stress phenotyping in plants. *Trends Plant Sci.* **2016**, *21*, 110–124. [CrossRef] [PubMed]
113. Naik, H.S.; Zhang, J.P.; Lofquist, A.; Assefa, T.; Sarkar, S.; Ackerman, D.; Singh, A.; Singh, A.K.; Ganapathysubramanian, B. A real-time phenotyping framework using machine learning for plant stress severity rating in soybean. *Plant Methods* **2017**, *13*, 23. [CrossRef] [PubMed]
114. Ghosal, S.; Blystone, D.; Singh, A.K.; Ganapathysubramanian, B.; Singh, A.; Sarkar, S. An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 201716999. [CrossRef] [PubMed]
115. Ubbens, J.R.; Stavness, I. Deep Plant Phenomics: A deep learning platform for complex plant phenotyping tasks. *Front. Plant Sci.* **2017**, *8*, 1190. [CrossRef] [PubMed]
116. Heckmann, D.; Schluter, U.; Weber, A.P.M. Machine learning techniques for predicting crop photosynthetic capacity from leaf reflectance spectra. *Mol. Plant* **2017**, *10*, 878–890. [CrossRef] [PubMed]
117. Ma, C.; Zhang, H.H.; Wang, X. Machine learning for Big Data analytics in plants. *Trends Plant Sci.* **2014**, *19*, 798–808. [CrossRef] [PubMed]
118. Brenchley, R.; Spannagl, M.; Pfeifer, M.; Barker, G.L.; D'Amore, R.; Allen, A.M.; McKenzie, N.; Kramer, M.; Kerhornou, A.; Bolser, D.; et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **2012**, *491*, 705–710. [CrossRef] [PubMed]
119. Mapleson, D.; Venturini, L.; Kaithakottil, G.; Swarbreck, D. Efficient and accurate detection of splice junctions from RNAseq with Portcullis. *bioRxiv* **2017**, 217620. [CrossRef]
120. Hecker, M.; Lambeck, S.; Toepfer, S.; Van Someren, E.; Guthke, R. Gene regulatory network inference: Data integration in dynamic models—A review. *Biosystems* **2009**, *96*, 86–103. [CrossRef] [PubMed]
121. Marbach, D.; Roy, S.; Ay, F.; Meyer, P.E.; Candeias, R.; Kahveci, T.; Bristow, C.A.; Kellis, M. Predictive regulatory models in *Drosophila melanogaster* by integrative inference of transcriptional networks. *Genome Res.* **2012**, *22*, 1334–1349. [CrossRef] [PubMed]
122. Rafalski, J.A. Novel genetic mapping tools in plants: SNPs and LD-based approaches. *Plant Sci.* **2002**, *162*, 329–333. [CrossRef]
123. Flint-Garcia, S.A.; Thornsberry, J.M.; Buckler, E.S. Structure of linkage disequilibrium in plants. *Annu. Rev. Plant Biol.* **2003**, *54*, 357–374. [CrossRef] [PubMed]
124. Buggs, R.J.A.; Renny-Byfield, S.; Chester, M.; Jordon-Thaden, I.E.; Viccini, L.F.; Chamala, S.; Leitch, A.R.; Schnable, P.S.; Barbazuk, W.B.; Soltis, P.S.; et al. Next-generation sequencing and genome evolution in allopolyploids. *Am. J. Bot.* **2012**, *99*, 372–382. [CrossRef] [PubMed]
125. Clevenger, J.P.; Korani, W.; Ozias-Akins, P.; Jackson, S. Haplotype-based genotyping in polyploids. *Front. Plant Sci.* **2018**, *9*, 564. [CrossRef] [PubMed]
126. Luo, R.; Sedlazeck, F.J.; Lam, T.-W.; Schatz, M. Clairvoyante: A multi-task convolutional deep neural network for variant calling in Single Molecule Sequencing. *bioRxiv* **2018**, 310458. [CrossRef]
127. Gottschalk, W.; Wolff, G. *Induced Mutations in Plant Breeding*; Springer: Berlin, Germany, 1983.
128. Jinek, M.; Chylinski, K.; Fonfara, I.; Hauer, M.; Doudna, J.A.; Charpentier, E. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science* **2012**, *337*, 816–821. [CrossRef] [PubMed]
129. Scheben, A.; Wolter, F.; Batley, J.; Puchta, H.; Edwards, D. Towards CRISPR/Cas crops—Bringing together genomics and genome editing. *New Phytol.* **2017**, *216*, 682–698. [CrossRef] [PubMed]

130. Scheben, A.; Edwards, D. Genome editors take on crops. *Science* **2017**, *355*, 1122–1123. [CrossRef] [PubMed]
131. Gao, C. The future of CRISPR technologies in agriculture. *Nat. Rev. Mol. Cell Biol.* **2018**, *5*, 275–276. [CrossRef] [PubMed]
132. Eeckhaut, T.; Lakshmanan, P.S.; Deryckere, D.; Van Bockstaele, E.; Van Huylenbroeck, J. Progress in plant protoplast research. *Planta* **2013**, *238*, 991–1003. [CrossRef] [PubMed]
133. Wolter, F.; Klemm, J.; Puchta, H. Efficient in planta gene targeting in Arabidopsis using egg-cell specific expression of the Cas9 nuclease of *Staphylococcus aureus*. *Plant J.* **2018**. [CrossRef] [PubMed]
134. Xu, H.; Xiao, T.F.; Chen, C.H.; Li, W.; Meyer, C.A.; Wu, Q.; Wu, D.; Cong, L.; Zhang, F.; Liu, J.S.; et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **2015**, *25*, 1147–1157. [CrossRef] [PubMed]
135. Chari, R.; Mali, P.; Moosburner, M.; Church, G.M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Meth.* **2015**, *12*, 823–826. [CrossRef] [PubMed]
136. Liang, G.; Zhang, H.M.; Lou, D.J.; Yu, D.Q. Selection of highly efficient sgRNAs for CRISPR/Cas9-based plant genome editing. *Sci. Rep.* **2016**, *6*, 21451. [CrossRef] [PubMed]
137. Lee, C.M.; Cradick, T.J.; Fine, E.J.; Bao, G. Nuclease target site selection for maximizing on-target activity and minimizing off-target effects in genome editing. *Mol. Ther.* **2016**, *24*, 475–487. [CrossRef] [PubMed]
138. Henry, V.J.; Bandrowski, A.E.; Pepin, A.S.; Gonzalez, B.J.; Desfeux, A. OMICtools: An informative directory for multi-omic data analysis. *Database* **2014**, *2014*, bau069. [CrossRef] [PubMed]
139. Liu, H.; Ding, Y.; Zhou, Y.; Jin, W.; Xie, K.; Chen, L.L. CRISPR-P 2.0: An improved CRISPR-Cas9 tool for genome editing in plants. *Mol. Plant* **2017**, *10*, 530–532. [CrossRef] [PubMed]
140. Xie, K.; Zhang, J.; Yang, Y. Genome-wide prediction of highly specific guide RNA spacers for CRISPR-Cas9-mediated genome editing in model plants and major crops. *Mol. Plant* **2014**, *7*, 923–926. [CrossRef] [PubMed]
141. Liu, H.; Wei, Z.; Dominguez, A.; Li, Y.; Wang, X.; Qi, L.S. CRISPR-ERA: A comprehensive design tool for CRISPR-mediated gene editing, repression and activation. *Bioinformatics* **2015**, *31*, 3676–3678. [CrossRef] [PubMed]
142. Doench, J.G.; Fusi, N.; Sullender, M.; Hegde, M.; Vaimberg, E.W.; Donovan, K.F.; Smith, I.; Tothova, Z.; Wilen, C.; Orchard, R.; et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **2016**, *34*, 184–191. [CrossRef] [PubMed]
143. Sattar, M.N.; Iqbal, Z.; Tahir, M.N.; Shahid, M.S.; Khurshid, M.; Al-Khateeb, A.A.; Al-Khateeb, S.A. CRISPR/Cas9: A practical approach in date palm genome editing. *Front. Plant Sci.* **2017**, *8*, 1469. [CrossRef] [PubMed]