SCIENCE CHINA PRESS

SCIENCE CHINA Life Sciences
life.scichina.com link.springer.com
https://doi.org/10.1007/s11427-024-2578-6

REVIEW

Computational tools for plant genomics and breeding

Hai Wang[1,2,3†*], Mengjiao Chen[4†], Xin Wei[5], Rui Xia[6], Dong Pei[4], Xuehui Huang[5] & Bin Han[7]

[1]State Key Laboratory of Maize Bio-breeding, Frontiers Science Center for Molecular Design Breeding, Joint International Research Laboratory of Crop Molecular Breeding, National Maize Improvement Center, College of Agronomy and Biotechnology, China Agricultural University, Beijing 100193, China;
[2]Sanya Institute of China Agricultural University, Sanya 572025, China;
[3]Hainan Yazhou Bay Seed Laboratory, Sanya 572025, China;
[4]State Key Laboratory of Tree Genetics and Breeding, Key Laboratory of Tree Breeding and Cultivation of the State Forestry and Grassland Administration, Research Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, China;
[5]Shanghai Key Laboratory of Plant Molecular Sciences, College of Life Sciences, Shanghai Normal University, Shanghai 200234, China;
[6]College of Horticulture, South China Agricultural University, Guangzhou 510640, China;
[7]National Center for Gene Research, CAS Center for Excellence in Molecular Plant Sciences, Chinese Academy of Sciences, Shanghai 200233, China

†These authors contributed equally
*Corresponding author (email: wanghai@cau.edu.cn)

Received 5 February 2024; Accepted 25 March 2024; Published online 23 April 2024

Plant genomics and crop breeding are at the intersection of biotechnology and information technology. Driven by a combination of high-throughput sequencing, molecular biology and data science, great advances have been made in omics technologies at every step along the central dogma, especially in genome assembling, genome annotation, epigenomic profiling, and transcriptome profiling. These advances further revolutionized three directions of development. One is genetic dissection of complex traits in crops, along with genomic prediction and selection. The second is comparative genomics and evolution, which open up new opportunities to depict the evolutionary constraints of biological sequences for deleterious variant discovery. The third direction is the development of deep learning approaches for the rational design of biological sequences, especially proteins, for synthetic biology. All three directions of development serve as the foundation for a new era of crop breeding where agronomic traits are enhanced by genome design.

bioinformatics | plant genomics | breeding by design

How we formulate and investigate scientific questions in plant genomics has been revolutionized in the past two decades: data-driven research is now at least as important as hypothesis-driven research. This trend was brought about by an exponential increase in the volume of data that transformed plant biology from a data-poor to a data-rich discipline. These datasets may take diverse forms, including genome sequences, epi-genetic marks, open chromatin, cistrome, 3D genome organization, transcriptome, proteome, metabolome, phenome, as well as their interactome. Data can also be generated from various genotypes, environmental conditions, tissue types, or even cells with different identities. All these boosted the development of numerous quantitative and predictive computational tools that help us to explore, summarize and visualize data, and to form new insightful, informative, and falsifiable hypotheses.

The way we design and engineer biological systems in model plants and crops has also been reformed in the past two decades. In plant genomics, high-quality reference genomes and pan-genomes, along with high-throughput genotyping and phenotyping, enabled unprecedented precision and power in the genetic dissection of complex traits by association mapping or linkage mapping. This enabled the development of cutting-edge statistical methods in genomic prediction and selection for crop breeding. In comparative genomics, efficient tools have been developed to align genomes within or between species, making it possible to depict in more detail the retention and fractionation of transposable elements and gene families after polyploidization. Comparative genomics also enabled the delineation of the evolutionary constraints in both coding and non-coding regions and the prediction of deleterious variants in constrained genomic loci. In synthetic biology, deep learning has revolutionized protein sequence-to-structure prediction and end-to-end structure-to-sequence protein design.

Numerous computational tools have been developed to tackle the above-mentioned questions, but a comprehensive review is still lacking to summarize their design principles and applications. In this review, we review widely-used computational tools and pipelines in plant genomics and crop breeding (Figure 1; Table S1 in Supporting Information). Current challenges and future perspectives in data mining for plant genomics and crop breeding are also discussed.

Genome assembly, annotation, and comparative genomics

Contig assembly

Before the advent of long-read sequencing platforms from Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT), genomes were assembled from short reads. For both types of sequencing platforms, it is generally recommended to perform quality trimming and checking on raw sequencing data using software such as Fastp (Chen et al., 2018) before genome assembly. Several assemblers are available. ALLPaths-LG is a

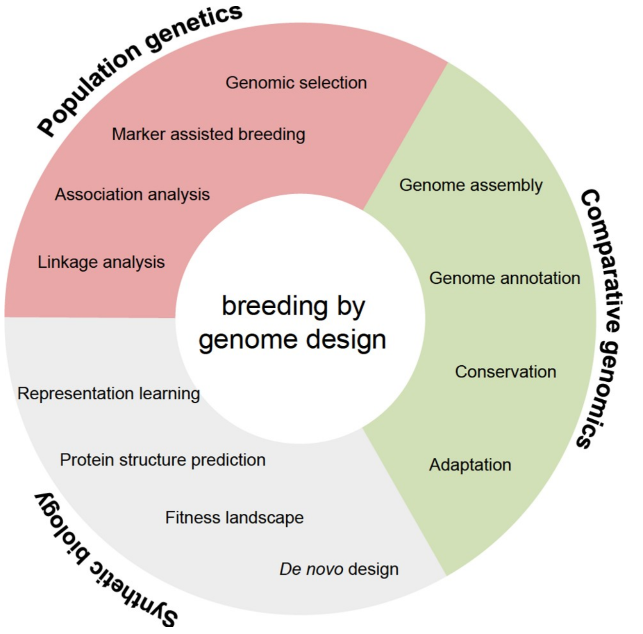

Figure 1. A summary of major branches in the field of plant genomics and breeding.

whole-genome shotgun assembler that can generate high-quality genome assemblies using short reads (Gnerre et al., 2011). It requires at least a "short library" with an average separation size slightly less than twice the read size and a "long library" with an insert size of approximately 3,000 base pairs. SOAPdenovo is a *de novo* draft assembly method specially designed to assemble Illumina GA short reads for human-sized genomes in a cost-effective way (Luo et al., 2012). SPAdes is used to assemble mitochondria and plastid genome sequences. SPAdes works with Illumina or IonTorrent reads and is capable of providing hybrid assemblies using PacBio, ONT and Sanger reads (Prjibelski et al., 2020). Minia is a short-read assembler based on a de Bruijn graph, capable of assembling a human-sized genome on a desktop in a day (Salikhov et al., 2014).

In recent years, short-read platforms have been gradually replaced by long-read technologies in genome assembly. Reads ranging from kilobases to megabases enable better resolution of structural variations and long repeats. ONT and PacBio are two major providers of commercial long-read sequencing technology. PacBio HiFi reads are characterized by lower base errors (currently <1%) but shorter lengths (10–25 kb), compared with ONT (10–100 kb, 1%–13% error rates). Therefore, long read assemblers require special algorithms and data structures that can accommodate errors in the sequencing data. Canu can reliably assemble near-complete eukaryotic chromosomes using either PacBio or Oxford Nanopore technologies. It adopted new overlapping and assembly algorithms to avoid collapsing diverged repeats and haplotypes (Koren et al., 2018; Koren et al., 2017; Nurk et al., 2020). Similar to Canu, NextDenovo is another string graph-based *de novo* assembler for long reads that uses a "correct-then-assemble" strategy, but it consumes less computing resources and storage (Hu et al., 2023). Falcon is a set of tools for fast aligning long reads for consensus and assembly. It takes PacBio long reads as input and outputs a fasta file of all of the primary contigs (Chin et al., 2013). Flye is another *de novo* assembler for single molecule sequencing reads, and it was primarily developed to run on uncorrected reads (Kolmogorov et al., 2019).

Before the highly accurate PacBio HiFi sequencing method, assembled contigs should be polished by NGS short reads to fix base errors. NextPolish fixes base errors (SNV/Indel) in the genome generated by noisy long reads (Hu et al., 2020). Pilon is a fully automated tool for correcting draft assemblies, e.g. correcting bases, fixing mis-assemblies, and filling gaps (Walker et al., 2014). Racon (standing for Rapid Consensus) is a standalone and fast consensus module to correct raw contigs generated by rapid assembly methods which do not include a consensus step (Vaser et al., 2017).

Scaffolding

Several scaffolding tools have been developed to construct chromosome-scale scaffolds from contigs. The positions and orientations of contigs can be determined *de novo* by using Hi-C data, or inferred from other phylogenetically closely related genomes based on synteny. The former method has been implemented in ALLHiC (Zhang et al., 2019) and 3D-DNA (Dudchenko et al., 2017), and the latter method is exemplified by RagTag, a collection of software tools for scaffolding and improving genome assemblies. RagTag accomplishes homology-based misassembly correction, homology-based assembly scaffolding, and scaffold merging (Alonge et al., 2022).

Quality checking

In addition to technical measures like contig N50 or scaffold N50, several tools were developed to assess the completeness of a genome assembly. Benchmarking Universal Single-Copy Orthologs (BUSCO) tests the presence of a set of single-copy orthologs conserved across a wide range of taxa (Simao et al., 2015). Compleasm is a faster and more accurate lightweight reimplementation of BUSCO (Huang and Li, 2023). It assesses assembly completeness for a human genome within 30 min and achieves improved agreement with annotation completeness. However, it is conceivable that adaptive loss of genes in specific species may lead to an under-estimation of the completeness of its genome. While BUSCO only assesses gene space, the LTR assembly index (LAI) is used to estimate genome completeness in repetitive genome regions by calculating the intactness of LTR retrotransposons (Ou et al., 2018). Another useful tool is Merqury, which achieves reference-free assembly evaluation by comparing *k*-mers in a *de novo* assembly to those in unassembled reads (Rhie et al., 2020).

Annotation of repetitive sequences and transposable elements

Annotation of repetitive sequences often precedes the annotation of genes to reduce the computational cost and enhance the accuracy of gene annotation. Tandem repeats, interspersed repeats, and low complexity DNA sequences can be identified by Tandem Repeats Finder (TRF) (Benson, 1999) and RepeatMasker (Chen, 2004). Transposable elements can be annotated by several general repeat annotators including RepeatModeler, Red, Generic Repeat Finder (GRF), and a repeat database Repbase. Among them, RepeatModeler and Repbase provide classification of TEs. For LTR retrotransposon identification, in addition to RepeatModeler and Repbase, structure-based methods for *de novo* LTR identification are also available, including LTR_STRUC, LTR_FINDER (Xu and Wang, 2007), LTRharvest, MGEScan3, LTR_retriever, and LtrDetector. For the identification of TIR transposons, especially MITEs, available tools include IRF, TIRvish, TIR-Learner, MITE-Hunter, detectMITE, MUSTv2, miteFinderII, and MITE-Tracker. However, Non-LTR retrotransposons (such as LINEs and SINEs) and helitrons lack terminal repeats, making their identification particularly challenging. Recently, a pipeline called Extensive *de novo* TE Annotator (EDTA) has been developed to combine several tools (LTRharvest, LTR_FINDER, LTR_retriever, GRF, TIR-Learner, HelitronScanner, and RepeatModeler) for *de novo* identification of each TE subclass, and it compiles the results into a comprehensive nonredundant TE library (Ou et al., 2019).

Annotation of coding genes

Determining the structure of protein-coding genes for a newly assembled genome is still a daunting task today and often involves incorporating evidence from multiple sources. A quick and inexpensive way of gene structure annotation is *ab initio* gene prediction. AUGUSTUS is a software for gene prediction in eukaryotic genomic sequences based on a generalized Hidden Markov Model. AUGUSTUS allows the user to impose constraints on the predicted gene structure (i.e. splice sites, translation initiation sites, stop codons, known exonic intervals, or intronic genomic intervals) and provides the most likely gene structure that complies with user-defined constraints (Stanke et al., 2006). Semi-HMM-based Nucleic Acid Parser (SNAP) is another *ab initio* gene finding program (Korf, 2004). SNAP is similar to other generalized hidden Markov model gene finders but is more easily adaptable to diverse organisms. It should be noted that sequence features such as codon bias and splicing signals vary from organism to organism, *ab initio* gene finders should be trained on the species being annotated or other closely related species to ensure high accuracy. GeneMark is a family of gene prediction programs (Lomsadze et al., 2005). For gene prediction in eukaryotes, genomes can be analyzed by the self-training GeneMark-ES. FGENESH is another hidden Markov model-based gene structure predictor for automatic prediction of genes in eukaryotic genomes, with a quality similar to manual annotation (Solovyev et al., 2006).

Transcriptome sequencing represents an important source of evidence for gene structure annotation. Trinity is a popular tool for *de novo* reconstruction of transcriptomes from illumine RNA sequencing (RNA-seq) data without a reference genome (Grabherr et al., 2011), and the resulting transcriptome can be mapped to the reference genome to guide gene structure annotation. StringTie can be used to update gene annotation by assembling RNA-Seq alignments into potential transcripts (Pertea et al., 2015). Similar to StringTie, Cuffliks is another tool for transcript assembly and is often used in the identification of unannotated transcripts (Trapnell et al., 2010).

Protein sequences from the species being sequenced or other closely related species are another important source of evidence for gene annotation. Exonerate is a general-purpose tool for pairwise sequence comparison, and can be used to map protein sequences to a genome (Slater and Birney, 2005). Gene Model Mapper (GeMoMa) infers protein-coding genes in a target genome from their homologs in an already-annotated reference genome. It also predicts splice sites by incorporating RNA-seq evidence (Keilwagen et al., 2016; Keilwagen et al., 2018). Another choice is GenomeThreader, which predicts gene structure by spliced alignments of cDNA/EST and/or protein sequences to the target genome (Gremme et al., 2005).

To generate the final report of gene structure annotation as a weighted consensus of all available evidence, EVidenceModeler (EVM) was developed to integrate results from *ab initio* gene predictions, protein alignments, and transcript alignments (Haas et al., 2008). EVM is often used with the Program to Assemble Spliced Alignments (PASA) (Haas et al., 2003) to yield a comprehensive and configurable annotation system that predicts protein-coding genes and alternatively spliced isoforms. MAKER is a genome annotation pipeline that integrates many tools with different functions, including identification of repeats, alignment of transcripts and proteins to a genome, *ab initio* gene prediction, and quality assessment of evidence. Moreover, higher quality gene models can be generated by retraining the gene prediction algorithm on the outputs of preliminary runs (Cantarel et al., 2008).

It should be noted that precise gene structure annotation is still a challenge today, and even the most state-of-the-art tool inevitably generates many errors. Therefore, manual correction of gene structure annotation is necessary. Recently the Arabidopsis community has been trying to polish the Arabidopsis annotation using the tool Apollo (Dunn et al., 2019). IGV-GSAman is another user-friendly tool for the manual edition of gene structure annotation (https://gitee.com/CJchen/IGV-sRNA).

Functional annotations of genes

Ideally, the function of a gene should be annotated solely by experimental evidence, such as the molecular and physiological consequences of genetic perturbation in this gene, its expression pattern, genetic interaction partners, direct or indirect physical interaction partners of its protein, as well as the subcellular localization and biochemical activity of its protein. In reality, as the functions of most genes in even model species have not been experimentally defined, functional annotations of genes are mostly obtained by transferring annotation from homologous proteins with similar domain architectures. InterPro (Mulder et al., 2007), Pfam (now part of InterPro) (Bateman et al., 2000), and PROSITE are popular databases of protein domains and families, coupled with online and stand-alone tools to predict domains and important sites in proteins submitted by the user, i.e. InterProScan (Quevillon et al., 2005), PfamScan (Bateman et al., 2000), and ScanProsite (de Castro et al., 2006). Although transfer of functional annotation has been considered most effective for closely-related proteins, this can actually be expanded to remotely-related proteins. Two recent studies used protein structure prediction coupled with structural similarity search to discover novel deaminases (Huang et al., 2023) and CRISPR-Cas systems (Altae-Tran et al., 2023).

The most widely adopted standardized gene annotation system is gene ontology (GO). GO describes our knowledge of the biological domain to molecular function, cellular components, and biological processes (Blake and Harris, 2002). GO terms are structured as a directed acyclic graph, and each term could have its parent term and child terms. Several homology-based GO annotation pipelines have been developed, including Blast2GO (Conesa et al., 2005), DAVID (Huang et al., 2007), InterPro2GO (Burge et al., 2012), and AgriGO (Du et al., 2010; Tian et al., 2017). MapMan (Usadel et al., 2009) and PageMan (Usadel et al., 2006) provide a set of ontology terms similar to GO terms, along with tools for automatic functional annotation and enrichment analysis. In addition, Kyoto Encyclopedia of Genes and Genomes (KEGG) is another well-known database resource for mapping genes to possible biological pathways (Kanehisa and Goto, 2000).

Comparative genomics

Comparative genomics is a field of biological research in which the genomic features of different organisms are compared with reveal their similarities, differences, evolutionary relationships, and underlying evolutionary forces. It also shows the history of chromosome structural rearrangements, and gene family expansion/contraction that might contribute to the adaptation of a taxa.

One important subject of comparative genomics is synteny, or collinearity. It refers to the preservation of the order of genes on a chromosome passed down from a common ancestor. Exceptional conservation of synteny can reflect important functional relationships between neighboring genes, and shared synteny is one of the most reliable criteria for establishing the orthology of genomic regions in different species. Several methods have been developed to identify conserved synteny blocks on chromosomes in deeply diverged eukaryotes. Commonly, dynamic programming was used to build chains of pairwise collinear genes, as in MCscan (Tang et al., 2008), MCScanX (Wang et al., 2012), and JCVI.

Widespread synteny within a single genome is often an indication of genome duplication. In evolutionary biology, whole genome duplication (WGD) or polyploidy is a fundamental driving force for the origin and diversification of organisms. Common approaches to detecting signals of ancient WGDs include the distribution of Ks and intragenomic collinearity. Relevant tools include WGDI (Sun et al., 2022), GenoDup (Mao, 2019), WGDdetector (Yang et al., 2019), and WGD (Zwaenepoel and Van de Peer, 2019). WGDdetector can be reliably applied to poor-quality genomes or transcriptomes, while WGD is designed specifically for the detection of ancient WGDs.

Identifying orthology between genes is often the first step toward a holistic understanding of evolution and diversity of gene families, and the extrapolation of biological knowledge among organisms. The most widely used methods for orthology inference can be classified into two distinct groups. The first strategy first infers pairwise relationships between genes from two species, and then extends orthology to multiple species by identifying ortholog groups spanning these species. Examples include MultiParanoid (Alexeyenko et al., 2006) and OMA (Altenhoff et al., 2011). The second strategy attempts to identify complete orthogroups, as in OrthoMCL (Li et al., 2003) and OrthoFinder (Emms and Kelly, 2015). Frequently, there was not a one-to-one correspondence in orthology between species due to gene family expansion/contraction, and such changes in gene family size can be identified using Computational Analysis of gene Family Evolution (CAFE) (Mendes et al., 2021).

In addition to gene order and gene content, the evolution of non-coding sequences is also an important topic in comparative genomics. Fast and sensitive tools have been developed to align very long DNA sequences at chromosome or even genome scale, facilitating direct comparison of two genomes at single nucleotide resolution. Such tools include Minimap2 (Li, 2018; Li, 2021), MUMmer (Marcais et al., 2018) and AnchorWave (Song et al., 2022). More specifically, cis-regulatory sequences of orthologs and paralogs can be compared with identified conserved noncoding sequences (CNS), which is an indicator of purifying selection and functional constraint. Popular tools designed for plants include STAG-CNS (Lai et al., 2017) and dCNS (Song et al., 2021).

Epigenetics and transcriptome profiling

Functional annotation of regulatory genomic regions

Numerous omics approaches have been developed to profile epigenetic marks, cistrome, and chromatin states to functionally annotate non-coding regions of the genome. This strategy has been extensively applied to the human and mouse genome in the ENCODE project, and has also been successfully used in some crop species. Cytosine methylation of DNA is an important epigenetic modification to control gene expression and genomic imprinting. Cytosine methylation can be detected by sequencing sodium bisulfite-treated DNA in a process termed bisulfite sequencing (BS-Seq). Bisulfite treatment of DNA converts non-methylated cytosines into uracils which are further converted into thymines by PCR amplification. Therefore, a mapper of bisulfite-treated sequences should be capable of coping with both natural variants and mismatches introduced by bisulfite treatment. Commonly used bisulfite read aligners include BatMeth2 (Zhou et al., 2019), BSMAP (Xi and Li, 2009), Bismark (Krueger and Andrews, 2011), BS-Seeker2 (Guo et al., 2013), BWA-meth, and BISulfite-seq CUI Toolkit (BISCUIT). The differential methylation analysis of BS-seq data can be achieved by Fisher's exact test, BSmooth, methylKit, methylSig, DSS, metilene, RADMeth, or Biseq.

Chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq), DNA Affinity Purification Sequencing (DAP-Seq), Cleavage Under Targets and Tagmentation (CUT&Tag), and Assay for Transposase Accessible Chromatin with high-throughput Sequencing (ATAC-Seq), and similar techniques has been extensively used to determine the chromatin state such as binding sites of chromatin associated proteins, histone modifications, and chromatin accessibility. In general, the first step of analysis starts with mapping raw reads to the reference genome by short read mappers such as bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2019) and bwa-mem (Li and Durbin, 2009). Reads are then mapped to the reference genome, and regions with significant enrichment of read coverage to the background are identified. Commonly used tools include MACS (Zhang et al., 2008), F-Seq2 (Zhao and Boyle, 2021), SICER, and RSEG. Differential analysis based on pre-defined peaks can be done with MACS and SICER, or more specific tools such as DiffBind and DBChIP, which adopted the statistical models of computational tools developed for differential expression analysis of RNA-seq data such as edgeR and DESeq. In recent years, similar techniques can also be carried out at single cell level and analyzed by tools such as ArchR (Granja et al., 2021), SnapATAC (Fang et al., 2021), and Signac (Stuart et al., 2021).

mRNA profiling

RNA-seq has become an indispensable tool for RNA biology, including transcriptome-wide analysis of differential gene expression, differential splicing of mRNAs, spatial-temporal control of gene expression, single-cell gene expression, and spatial transcriptomics. RNA molecules are most commonly sequenced by short-read techniques, but long-read platforms have become more popular in recent years due to their ability to better identify different isoforms resulting from alternative splicing. Moreover, nanopore direct RNA sequencing makes it possible to read full-length sequences and modifications on the RNA molecules simultaneously. After sequencing, reads were quality trimmed by Trimmomatic (Bolger et al., 2014) or sickle (https://github.com/najoshi/sickle), and then mapped to a reference genome. Mapping of RNA-Seq reads is fundamentally different from that of re-sequencing in that splice aware aligners should be used to properly map reads spanning splice junctions. TopHat2 is a fast spliced aligner for RNA-sequence experiments (Kim et al., 2013), and it has now been superseded by HISAT2, which is more efficient and accurate and provides the same core functionality (Kim et al., 2019). Spliced Transcripts Alignment to a Reference (STAR) is another ultra-fast RNA-seq read mapper that supports both splice-junction and fusion read detection (Dobin et al., 2013). Comprehensive benchmarking of RNA-Seq aligners for short reads (Baruzzo et al., 2017) and long reads (Krizanovic et al., 2018) indicates the importance of choosing the right tool and optimized parameters to achieve optimal accuracy in read mapping.

RNA-Seq alignments in the SAM format produced by splice aware aligners are then converted to BAM files and sorted by coordinates using SAMtools, and subjected to gene-level or transcript-level expression quantification. StringTie, in addition to its major role as an assembler of transcripts, also provides estimates of normalized expression levels in FPKM or TPM units (Pertea et al., 2015). FeatureCounts from the SourceForge Subread package is a highly efficient read summarization program that counts mapped reads for any genomic features specified by the user, such as genes, exons, promoters, gene bodies, and genomic bins (Liao et al., 2014). FeatureCounts is compatible with some downstream analysis such as the identification of differentially expressed genes by DESeq2, which requires raw counts of RNA-Seq reads. More recently, k-mer-based alignment-free transcript quantification pipelines, such as Kallisto (Bray et al., 2016) and Salmon (Patro et al., 2017), were developed. The pseudoalignments of Kallisto and quasi-mapping along with GC- and sequence-bias corrections of Salmon make them orders of magnitude faster than alignment-based pipelines.

Non-coding RNA profiling

Next generation sequencing also provides new opportunities for non-coding RNA profiling. Non-coding RNA (ncRNA) PROfiling in small RNA (sRNA)-seq (ncPRO-seq) is a pipeline that performs detailed profiling analysis on sRNAs. The pipeline is flexible as it can be applied on sRNAs derived from annotated non-coding regions in miRBase, Rfam and RepeatMasker, or user-defined genomic regions (Chen et al., 2012). Some tools have been developed for circular RNA discovery. The CIRCexplorer pipeline can systematically annotate different types of alternative back-splicing and alternative splicing events in circRNAs from various cell lines, and evaluate circular and linear RNA expression individually (Ma et al., 2019; Zhang et al., 2016). Circular RNA detection and quantification can also be achieved from RNA-seq reads by Find_circ (Memczak et al., 2013) and CIRIquant (Zhang et al., 2020).

Differential expression, co-expression, and regulatory networks

Quantification of RNA molecules is in most cases only the first step of a project. In comparative RNA-Seq assays, a fundamental task is the analysis of differentially expressed genes across experimental conditions. To this end, DESeq2 (Love et al., 2014), Cuffdiff, and EdgeR (Robinson et al., 2010) were developed. All three tools assume a negative binomial distribution of read counts with generalized linear models to determine the significance of differential expression, to cope with difficulties such as small replicate numbers, discreteness, large dynamic range, and the presence of outliers. When sample sizes are large, the Wilcoxon rank-sum test was recommended for its better performance in the control of false discovery rate (Li et al., 2022).

Co-expression, genetic regulatory networks (GRNs), and enrichment of cis-regulatory motifs are also frequently analyzed. Weighted gene co-expression network analysis (WGCNA) allows the convenient construction of a co-expression network from transcriptome quantification datasets. It also facilitates the identification of modules, intramodular hubs, and the comparison of the topology of different networks (Langfelder and Horvath, 2008). GENIE3 aims to construct a GRN by predicting the expression pattern of a target gene from the expression patterns of all the other input genes using random forests or extra trees (Huynh-Thu et al., 2010). Genes responsive to specific exogenous or endogenous cues at the transcriptional level often share common regulators and thus common *cis*-regulatory elements. The MEME Suite allows the discovery of novel enriched motifs in co-regulated genes (Bailey et al., 2015).

Plant breeding by design

Genotyping

In genotyping, NGS short-read platforms are still preferred over long-read techniques due to their lower cost. After re-sequencing, low-quality nucleotides, adaptors, and barcodes are trimmed by tools such as Trimmomatic (Bolger et al., 2014) or Sickle (https://github.com/najoshi/sickle), followed by quality checking using FastQC. Reads are then mapped to a reference genome. The most popular tools include bwa-mem (Li and Durbin, 2009) and bowtie2 (Langmead and Salzberg, 2012; Langmead et al., 2019). Recently, a beta version of bwa-mem2 has been released, and it is about twice as fast as bwa-mem without a penalty in accuracy (Vasimuddin et al., 2019). Minimap2 was developed for PacBio and Nanopore long-read alignment (Li, 2018; Li, 2021). The resulting alignment files in SAM format need to be converted to BAM format and sorted by position using SAMtools (Danecek et al., 2021) or SAMbamba (Tarasov et al., 2015).

Genome Analysis Toolkit (GATK) has long been the industry standard for identifying SNPs and InDels from the above-mentioned BAM files (McKenna et al., 2010). But for plants with extremely large and complex genomes (such as wheat), it remains challenging to use GATK, and BCFtools is more preferred in such cases. Recently, convolutional deep neural networks have been adopted in variant calling with state-of-the-art accuracies, such as DeepVariant (Poplin et al., 2018) and Clairvoyante (Luo et al., 2019). The identified genomic variants are in Variant Call Format (VCF) or its binary counterpart BCF, and they can be conveniently processed by Picard tools, VCFtools, or BCFtools (Danecek et al., 2011).

In recent years, functional mapping of structural variants (SVs) has become more popular due to the dramatic decrease in the cost of long-read sequencing techniques. SVs are generally defined as large genomic alterations longer than 50 bp, including copy number variants (CNVs), insertions, inversions, translocations, repetitive sequence expansions, and their complex combinations. SVs are nowadays commonly detected by long-read sequencing platforms and can be detected either by assembling reads for comparison between genomes or by read alignment to a reference genome. Alignment-based SV callers include software tools such as cuteSV, Duet, NanoVar, SVIM, Picky, and Sniffles2 (Smolka et al., 2024). Assembly-based SV calling methods identify SVs by assembly of contigs followed by mapping of contigs to a reference genome. Representative tools include SVIM-asm (Heller and Vingron, 2021), PAV, Assemblytics and SyRI (Goel et al., 2019).

After variant calling, each sample in the population always contains missing genotypes due to limited sequencing depth. As most downstream analyses (such as association analysis and genomic prediction) do not accept missing genotype data, genotype imputation is needed to infer the genotypes for unobserved variant sites based on their linkage disequilibrium with observed markers. It also enhances the statistical power and resolution of GWAS and genomic prediction. Imputation often involves a phasing step, which refers to the statistical estimation of haplotypes from genotype data. Most phasing tools nowadays are based on the Li and Stephens Hidden Markov Model (HMM), including fastPHASE (Scheet and Stephens, 2006), BEAGLE (Browning et al., 2018), IMPUTE (Howie et al., 2012), and SHAPEIT (Delaneau et al., 2011). Among them, FastPHASE, BEAGLE, and IMPUTE also provide tools for genotype imputation. Recently, OutcrossSeq was developed for genotyping and imputation in the recombinant populations of outcrossing plants, based on whole-genome low-coverage resequencing data (Chen et al., 2021).

It is also worth mentioning that in plants, a single reference genome is often not adequate to fully capture all genomic variants in populations due to tremendous genomic diversity. Therefore, pan-genomes constructed by comparing high quality genome assemblies or graph-based approaches are becoming more popular in variant calling. To better leverage the diversity captured by plant pan-genomes, the Practical Haplotype Graph (PHG) was developed to impute complete genomes from low density sequence or variant data (Bradbury et al., 2022). Genetic mapping on a pan-genome is helpful to map more quantitative trait loci and alleviate "the lost heritability" problem.

Genetic mapping of quantitative trait loci

The genetic architecture of complex traits, as revealed by association or linkage analysis, is still a field of active research due to its fundamental role in crop genetic improvement. In GWAS, phenotype and genotype data are collected from a carefully chosen diversity panel, and the markers significantly associated with the phenotype are identified by statistical methods. Some GWAS software packages are available, such as GEMMA (Zhou and Stephens, 2012), TASSEL (Bradbury et al., 2007), and GAPIT (Lipka et al., 2012; Tang et al., 2016; Wang and Zhang, 2021).

The earliest GWAS was performed in a naive fashion by performing a statistical test for each variation with the null hypothesis that there is no difference in the phenotype between genotype groups, followed by multiple testing corrections. However, such a procedure suffers from high rates of false positives resulting from population stratification and relatedness. To overcome this problem, the general linear model (GLM) was invented by taking the population structure as cofactors. The population structure can be represented either as subpopulation membership calculated by STRUCTURE (Pritchard et al., 2000) or principal components (Price et al., 2006). As a further improvement to GLM, the mixed linear model (MLM) considers both a fixed effect of population structure and a random effect of kinship (i.e. relatedness among all individuals) (Yu et al., 2006).

Several methods have been developed to calculate MLM equations faster, including efficient mixed-model association (EMMA) (Kang et al., 2008), population parameters previously determined (P3D) (Zhang et al., 2010), EMMA eXpedited (EMMAX) (Kang et al., 2010), factored spectrally transformed linear mixed models (FaST-LMM) (Lippert et al., 2011), and genome-wide efficient mixed model analysis (GEMMA) (Zhou and Stephens, 2012). More recently, GWAS of biobank-scale data was made possible by fastGWA, which controls population stratification by principal components, and relatedness by a sparse genetic relationship matrix (Jiang et al., 2019). Methods were also developed to improve the power of GWAS. The compressed MLM (CMLM) (Zhang et al., 2010) and enriched CMLM (ECMLM) (Li et al., 2014) improve the power by using a lower-rank kinship matrix. The power of GWAS can also be improved by incorporating multiple markers in the association model simultaneously, as implemented in the multi-locus mixed model (MLMM) (Segura et al., 2012), fixed and random model circulating probability unification (FarmCPU) (Liu et al., 2016), and Bayesian information and LD iteratively nested keyway (BLINK) (Huang et al., 2019). Although the most widely used genotype data are still SNPs and short InDels, SVs such as CNVs and presence/absence variations (PAVs) have attracted more attention in recent years. Moreover, *k*-mer-based GWAS was developed to detect QTLs not present in the reference genome (Voichek and Weigel, 2020). Although GWAS associates genomic variants directly to terminal traits, association analysis can also be conducted between genomic variants and endophenotypes (such as gene expression level, epigenetic marks, proteome, and metabolites), or between endophenotypes and terminal traits.

In crops, linkage analysis is still widely used, especially when the minor allele frequency of the causal variant is too low for GWAS. Various QTL mapping procedures have been proposed, such as composite interval mapping (CIM), multiple interval mapping with forward search (MIMF), multiple interval mapping with regression forward selection (MIMR), inclusive composite interval mapping (ICIM), multiple-QTL Mapping (MQM), and network interval mapping (NWIM). Most of these procedures have been implemented in a number of popular tools such as QTL Cartographer, IciMapping, MapQTL and QTLnetwork.

Although association and linkage analysis has become a routinely used technique, pinpointing the causal variant underlying a QTL is still a daunting task due to the linkage disequilibrium between the causal variant and neighboring variants. Computational tools that predict the effects of variants are often used in this scenario, such as variant effect predictor (VEP) (McLaren et al., 2016) and SnpEff (Cingolani et al., 2012), followed by wet lab experiments to assay the molecular or physiological effect of each candidate variant.

The breeding of crop ideotypes

Plant molecular breeding methods can be categorized into two primary groups, depending on their reliance on known quantitative trait loci (QTL) or genes. The first category employs genome-wide molecular markers to estimate the breeding value of individual genomes, a method commonly referred to as genomic selection breeding. This approach generally does not require a comprehensive genetic analysis of the target varieties. Instead, it relies on the provision of extensive genotype and phenotype data to training populations. Various models and software tools, including rrBLUP (Endelman, 2011), gBLUP (Clark and van der Werf, 2013), sommer (Covarrubias-Pazaran, 2016), SeqBreed (Pérez-Enciso et al., 2020), BGLR (Pérez-Rodríguez and de los Campos, 2022), ASREML (Butler et al., 2017), and CropGS-Hub (Chen et al., 2024), are utilized for predicting breeding values. In addition, deep learning methods have been adopted in genomic predictions, as exemplified by DNNGP (Wang et al., 2023) and DeepGS (Ma et al., 2018). The second category of breeding methods is grounded on an in-depth genetic analysis of plant agronomic traits. This approach requires breeders to possess a profound understanding of the target varieties and a familiarity with the characteristics of the identified QTLs or genes. Crop varieties suffering from inferior alleles of QTLs or genes can be improved through the replacement or modification of these disadvantageous alleles. This method is commonly known as molecular design breeding. In its early stages, this approach primarily made use of molecular markers linked to QTL genes for marker-assisted selection breeding. For instance, SSR markers, CAPS markers, and Indel markers were employed for assisted selection breeding. Software tools such as MISA (Beier et al., 2017), Primer3 (Untergasser et al., 2012), GATK (McKenna et al., 2010), SNP2CAPS (Thiel et al., 2004), and Dindel (Albers et al., 2011) were instrumental in marker design. However, there was a notable absence of software tools capable of guiding breeders in the overall design of their breeding programs. Additionally, the molecular markers used in marker-assisted selection did not affect gene function. They were just linked to quantitative trait nucleotides (QTNs) that determine gene function. This occasionally resulted in false positives during the breeding selection process. To address these challenges, researchers developed RiceNavi (Wei et al., 2021) and its web version (http://www.xhhuanglab.cn/tool/RiceNavi.html). RiceNavi enables the direct use of causal variants of QTL genes in rice molecular design breeding. By utilizing RiceNavi, customized improvements in rice agronomic traits can be achieved in less than three years, as exemplified by its application in the improvement of a widely cultivated indica rice variety. Conducting genotype simulations of breeding offspring with RiceNavi requires the use of PedigreeSim (Voorrips and Maliepaard, 2012).

Gene editing represents a novel molecular breeding technology developed in recent years. It not only enables the precise editing of inferior alleles but also facilitates the creation of new mutations in crop varieties that do not exist naturally. Breeders can utilize RiceNavi (Wei et al., 2021) to identify inferior alleles. Furthermore, software tools like CRISPR-GE (Xie et al., 2017) and CRISPRdirect (Naito et al., 2015) assist in the design of gene-editing primers. The screening of positive edited lines is conducted using software such as Hi-Tom (Liu et al., 2019).

Rational design of proteins

Two major objectives dominate current research in protein engineering: one is defining the fitness landscapes of proteins for disease diagnosis in humans and purging of deleterious variants in crop breeding, the other is designing proteins with specific biochemical properties to meet the demand of synthetic biology. Both objectives were backed by two sets of tools serving as infrastructures. The first tool set contains deep learning models that predict protein structures from sequences, such as AlphaFold (Jumper et al., 2021; Senior et al., 2020), transform-restrained Rosetta (trRosetta) (Yang et al., 2020), and RoseTTAFold (Baek et al., 2021). The second tool set comprises deep learning models that generate low dimensional latent space representations of protein sequences, often achieved by unsupervised learning of a huge repertoire of protein sequences. For example, UniRep was trained to perform the next amino-acid prediction and learn how to represent proteins as a fixed-length vector (Alley et al., 2019). A variational auto-encoder (VAE) model was trained to generate low dimensional latent space representations of protein sequences (Ding et al., 2019). In another study, representation learning of proteins was achieved by training a transformer model with the masked language modeling objective (Rives et al., 2021), such dimension reduction techniques greatly alleviate "the curse of dimensionality" in downstream sequence-function mapping.

The above-mentioned low dimensional representations of protein sequences can be used to model protein fitness landscapes and other properties such as protein mutational stability landscapes. DeepSequence predicts the effect of mutations in protein sequences by considering high-order dependencies in addition to site-specific constraints (Riesselman et al., 2018). ECNet (evolutionary context-integrated neural network) is a deep-learning algorithm that exploits evolutionary contexts to predict functional fitness for protein engineering (Luo et al., 2021). AlphaMissense, an adaptation of AlphaFold fine-tuned on human and primate variant population frequency databases, predicts missense variant pathogenicity (Cheng et al., 2023). It is conceivable that in the future, integration of machine learning with large-scale assays can better explore the fitness landscape of proteins. Moreover, deep learning models have been trained to use low dimensional representations of protein sequences or protein sequences themselves to predict diverse properties of proteins, such as protein-protein interactions (Homma et al., 2023; Wang et al., 2019), protein classifications (Strodthoff et al., 2020), protein functions (Gligorijevic et al., 2021), and GO annotations (Brandes et al., 2022).

Deep neural networks trained to predict protein structures from sequences can be inverted to design new proteins, as exemplified by deep network hallucination (Anishchenko et al., 2021). ProteinMPNN designs proteins by predicting a likely sequence that matches a predefined protein backbone (Dauparas et al., 2022). Similarly, constrained hallucination and protein inpainting were used to design proteins that match predefined functional residues (Wang et al., 2022). Indeed, various deep learning tools have been proven successful in designing proteins with specific functions, such as antimicrobial peptides (Das et al., 2021; Pandi et al., 2023), luciferase (Hawkins-Hooker et al., 2021), recombinase (Schmitt et al., 2022), antibody (Hie et al., 2024), transposase (Zhou et al., 2023), and proteinous binders of target proteins (Cao et al., 2022; Gainza et al., 2023; Torres et al., 2024; Yang et al., 2023).

Caveats and future perspectives

How to choose good scientific problems in plant genomics?

Choosing a good scientific problem is at the center of doing good science. Uri Alon reformulated this problem as equivalent to optimizing both feasibility (easy or hard to investigate) and interest (large or small gain in knowledge) according to one's current career stage (Alon, 2009). In plant genomics, what will be the next breakthrough and how to bring it about? Answering questions like this takes a vast body of knowledge and a deep understanding of biology. Fluency in programming and familiarity with existing computational tools by continuous practice is also crucial, as this allows one to test the hypotheses on large datasets efficiently. We also suggest data analysts get their feet wet at wet lab experiments so that they would have first-hand experience with the "reproducibility crisis" in molecular biology (Adam, 2023) to prevent "garbage in, garbage out" in data science.

How to develop your intuition for biological data analysis?

Scientific discoveries and formulation of hypotheses in plant genomics sometimes require intuition, which involves the ability to make educated guesses, see patterns, and make connections among different ideas. Intuition in plant genomics is often developed through practice, experience, and exposure to a wide variety of problems and concepts. For data analysts (especially novices), it is important to know how to view genomic data in order to have a visual feel of what the real data looks like. Tools for data browsing, like IGV (Thorvaldsdottir et al., 2013), JBrowse (Buels et al., 2016), and GBrowse (Stein, 2013), are often important starting points for finding new data patterns and bring forward new hypotheses and insights.

What should I do when having difficulties in installing a software?

Sometimes installing software on a Linux server can be frustrating, especially when you are a novice and not familiar with environment variables such as PATH and LIBRARY_-PATH. In addition, academic software tools are often delivered without their dependencies (let alone dependencies of dependencies), making their installation and maintenance extremely difficult. In such cases, the best practice would be to install them in isolated environments. Bioconda, a channel for the conda package manager, hosts numerous bioinformatics software that can be easily installed. Containers such as docker and singularity are also popular tools to find, use and share bioinformatics tools.

Can I trust a software without understanding its inner workings?

While it is generally unnecessary to reinvent the wheel, it is quite risky to put your full trust in a software. Sometimes extra efforts should be made to empirically benchmark various softwares and parameters on real or synthetic datasets, and to understand all the details under the hood by reading the source code. Otherwise, you may apply the software on unsuitable datasets with inappropriate parameters, or misinterpret the results. For example, DESeq2 assumes negative binomial distribution for counts distribution. You should probably resort to other statistical models if this prerequisite is not fulfilled. Besides, the association between a gene and a trait in a TWAS does not necessarily mean a causal relationship. Similarly, the co-expression of a transcription factor and another gene detected by WGCNA does not secure a regulatory relationship.

How to enhance the reusability of my data and reproducibility of my analysis?

To enhance the reusability of data and code, the findability, accessibility, interoperability, and reuse (FAIR) guiding principles for scientific data management and stewardship were proposed (Wilkinson et al., 2016). Metadata, data and code should be stored in readily accessible repositories such as Figshare, GitHub, or Bitbucket. The data needs to interoperate with applications or workflows for processing and analysis. Ideally, a workflow management system such as Snakemake can be used to create reproducible and scalable data analyses (Molder et al., 2021).

Democratization of biological data analysis

Is there a way for wet-lab researchers to analyze their data without investing too much of their time and effort in learning programming? In computer science, the low code or codeless programming movement is becoming popular. This is a digital philosophy that allows anyone to create applications and programs through "visual programming". A similar philosophy is also welcomed in the biology community. Take TBtools as an example, by incorporating over 130 functions, it is like a Swiss army knife for genomic data analysis. More importantly, it harbors a user-friendly interface, facilitating quick point-and-click data analysis (Chen et al., 2020).

Compliance and ethics

The author(s) declare that they have no conflict of interest.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (2022YFD1201100), the National Natural Science Foundation of China (32071464), Chinese Universities Scientific Fund (2023RC002), and the 2115 Talent Development Program of China Agricultural University.

Supporting information

The supporting information is available online at https://doi.org/10.1007/s11427-024-2578-6. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

Adam, D. (2023). What reproducibility crisis? New research protocol yields ultra-high replication rate. Nature 623, 467–468.

Albers, C.A., Lunter, G., MacArthur, D.G., McVean, G., Ouwehand, W.H., and Durbin, R. (2011). Dindel: accurate indel calls from short-read data. Genome Res 21, 961–973.

Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E.L.L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. Bioinformatics 22, e9–e15.

Alley, E.C., Khimulya, G., Biswas, S., AlQuraishi, M., and Church, G.M. (2019). Unified rational protein engineering with sequence-based deep representation learning. Nat Methods 16, 1315–1322.

Alon, U. (2009). How to choose a good scientific problem. Mol Cell 35, 726–728.

Alonge, M., Lebeigle, L., Kirsche, M., Jenike, K., Ou, S., Aganezov, S., Wang, X., Lippman, Z.B., Schatz, M.C., and Soyk, S. (2022). Automated assembly scaffolding using RagTag elevates a new tomato system for high-throughput genome editing. Genome Biol 23, 258.

Altae-Tran, H., Kannan, S., Suberski, A.J., Mears, K.S., Demircioglu, F.E., Moeller, L., Kocalar, S., Oshiro, R., Makarova, K.S., Macrae, R.K., et al. (2023). Uncovering the functional diversity of rare CRISPR-Cas systems with deep terascale clustering. Science 382, eadi1910.

Altenhoff, A.M., Schneider, A., Gonnet, G.H., and Dessimoz, C. (2011). OMA 2011: orthology inference among 1,000 complete genomes. Nucleic Acids Res 39, D289–D294.

Anishchenko, I., Pellock, S.J., Chidyausiku, T.M., Ramelot, T.A., Ovchinnikov, S., Hao, J., Bafna, K., Norn, C., Kang, A., Bera, A.K., et al. (2021). De novo protein design by deep network hallucination. Nature 600, 547–552.

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science 373, 871–876.

Bailey, T.L., Johnson, J., Grant, C.E., and Noble, W.S. (2015). The MEME suite. Nucleic Acids Res 43, W39–W49.

Baruzzo, G., Hayer, K.E., Kim, E.J., Di Camillo, B., FitzGerald, G.A., and Grant, G.R. (2017). Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods 14, 135–139.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L., and Sonnhammer, E.L. (2000). The Pfam protein families database. Nucleic Acids Res 28, 263–266.

Beier, S., Thiel, T., Münch, T., Scholz, U., and Mascher, M. (2017). MISA-web: a web server for microsatellite prediction. Bioinformatics 33, 2583–2585.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27, 573–580.

Blake, J.A., and Harris, M.A. (2002). The gene ontology (GO) project: structured vocabularies for molecular biology and their application to genome and expression analysis. Curr Protoc Bioinformatics doi: 10.1002/0471250953.bi0702s00.

Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120.

Bradbury, P.J., Casstevens, T., Jensen, S.E., Johnson, L.C., Miller, Z.R., Monier, B., Romay, M.C., Song, B., and Buckler, E.S. (2022). The Practical Haplotype Graph, a platform for storing and using pangenomes for imputation. Bioinformatics 38, 3698–3702.

Bradbury, P.J., Zhang, Z., Kroon, D.E., Casstevens, T.M., Ramdoss, Y., and Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23, 2633–2635.

Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., and Linial, M. (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics 38, 2102–2110.

Bray, N.L., Pimentel, H., Melsted, P., and Pachter, L. (2016). Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34, 525–527.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A one-penny imputed genome from next-generation reference panels. Am J Hum Genet 103, 338–348.

Buels, R., Yao, E., Diesh, C.M., Hayes, R.D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C.G., Lewis, S.E., Stein, L., et al. (2016). JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 17, 66.

Burge, S., Kelly, E., Lonsdale, D., Mutowo-Muellenet, P., McAnulla, C., Mitchell, A., Sangrador-Vegas, A., Yong, S.Y., Mulder, N., and Hunter, S. (2012). Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. Database 2012(0), bar068.

Butler, D., Cullis, B., Gilmour, A., Gogel, B., and Thompson, R. (2017). ASReml-R reference manual version 4. VSN International Ltd, Hemel Hempstead, UK.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sánchez Alvarado, A., and Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res 18, 188–196.

Cao, L., Coventry, B., Goreshnik, I., Huang, B., Sheffler, W., Park, J.S., Jude, K.M., Marković, I., Kadam, R.U., Verschueren, K.H.G., et al. (2022). Design of protein-binding proteins from the target structure alone. Nature 605, 551–560.

Chen, C., Chen, H., Zhang, Y., Thomas, H.R., Frank, M.H., He, Y., and Xia, R. (2020). TBtools: an integrative toolkit developed for interactive analyses of big biological data. Mol Plant 13, 1194–1202.

Chen, C.J., Servant, N., Toedling, J., Sarazin, A., Marchais, A., Duvernois-Berthet, E., Cognat, V., Colot, V., Voinnet, O., Heard, E., et al. (2012). ncPRO-seq: a tool for annotation and profiling of ncRNAs in sRNA-seq data. Bioinformatics 28, 3147–3149.

Chen, J., Tan, C., Zhu, M., Zhang, C., Wang, Z., Ni, X., Liu, Y., Wei, T., Wei, X.F., Fang, X., et al. (2024). CropGS-Hub: a comprehensive database of genotype and phenotype resources for genomic prediction in major crops. Nucleic Acids Res 52, D1519–D1529.

Chen, M., Fan, W., Ji, F., Hua, H., Liu, J., Yan, M., Ma, Q., Fan, J., Wang, Q., Zhang, S., et al. (2021). Genome-wide identification of agronomically important genes in outcrossing crops using OutcrossSeq. Mol Plant 14, 556–570.

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics 5.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34, i884–i890.

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L.H., Zielinski, M., Sargeant, T., et al. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science 381, eadg7492.

Chin, C.S., Alexander, D.H., Marks, P., Klammer, A.A., Drake, J., Heiner, C., Clum, A., Copeland, A., Huddleston, J., Eichler, E.E., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods 10, 563–569.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly 6, 80–92.

Clark, S.A., and van der Werf, J. (2013). Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. Methods Mol Biol 1019, 321–330.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21, 3674–3676.

Covarrubias-Pazaran, G. (2016). Genome-assisted prediction of quantitative traits using the R package sommer. PLoS ONE 11, e0156744.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics 27, 2156–2158.

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., et al. (2021). Twelve years of SAMtools and BCFtools. GigaScience 10, giab008.

Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., dos Santos, C., Chen, P.Y., et al. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nat Biomed Eng 5, 613–623.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R.J., Milles, L.F., Wicky, B I.M., Courbet, A., de Haas, R.J., Bethel, N., et al. (2022). Robust deep learning-based protein sequence design using ProteinMPNN. Science 378, 49–56.

de Castro, E., Sigrist, C.J.A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P.S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Res 34, W362–W365.

Delaneau, O., Marchini, J., and Zagury, J.F. (2011). A linear complexity phasing method for thousands of genomes. Nat Methods 9, 179–181.

Ding, X., Zou, Z., and Brooks Charles L., I. (2019). Deciphering protein evolution and fitness landscapes with latent space models. Nat Commun 10, 5644.

Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15–21.

Du, Z., Zhou, X., Ling, Y., Zhang, Z., and Su, Z. (2010). agriGO: a GO analysis toolkit for the agricultural community. Nucleic Acids Res 38, W64–W70.

Dudchenko, O., Batra, S.S., Omer, A.D., Nyquist, S.K., Hoeger, M., Durand, N.C., Shamim, M.S., Machol, I., Lander, E.S., Aiden, A.P., et al. (2017). De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. Science 356, 92–95.

Dunn, N.A., Unni, D.R., Diesh, C., Munoz-Torres, M., Harris, N.L., Yao, E., Rasche, H., Holmes, I.H., Elsik, C.G., and Lewis, S.E. (2019). Apollo: democratizing genome annotation. PLoS Comput Biol 15, e1006790.

Emms, D.M., and Kelly, S. (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol 16, 157.

Endelman, J.B. (2011). Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4, 250–255.

Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X., Xie, F., et al. (2021). Comprehensive analysis of single cell ATAC-seq data with SnapATAC. Nat Commun 12, 1337.

Gainza, P., Wehrle, S., Van Hall-Beauvais, A., Marchand, A., Scheck, A., Harteveld, Z., Buckley, S., Ni, D., Tan, S., Sverrisson, F., et al. (2023). De novo design of protein interactions with learned surface fingerprints. Nature 617, 176–184.

Gligorijevic, V., Renfrew, P.D., Kosciolek, T., Leman, J.K., Berenberg, D., Vatanen, T., Chandler, C., Taylor, B.C., Fisk, I.M., Vlamakis, H., et al. (2021). Structure-based protein function prediction using graph convolutional networks. Nat Commun 12, 3168.

Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108, 1513–1518.

Goel, M., Sun, H., Jiao, W.B., and Schneeberger, K. (2019). SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. Genome Biol 20, 277.

Gremme, G., Brendel, V., Sparks, M.E., and Kurtz, S. (2005). Engineering a software tool for gene structure prediction in higher organisms. Inf Software Tech 47, 965–978.

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. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29, 644–652.

Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet 53, 403–411.

Guo, W., Fiziev, P., Yan, W., Cokus, S., Sun, X., Zhang, M.Q., Chen, P.Y., and Pellegrini, M. (2013). BS-Seeker2: a versatile aligning pipeline for bisulfite sequencing data. BMC Genomics 14, 774.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith, R.K., Jr.Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D., Salzberg, S.L., and White, O. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res 31, 5654–5666.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C. R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol 9, R7.

Hawkins-Hooker, A., Depardieu, F., Baur, S., Couairon, G., Chen, A., and Bikard, D. (2021). Generating functional protein variants with variational autoencoders. PLoS Comput Biol 17, e1008736.

Heller, D., and Vingron, M. (2021). SVIM-asm: structural variant detection from haploid and diploid genome assemblies. Bioinformatics 36, 5519–5521.

Hie, B.L., Shanker, V.R., Xu, D., Bruun, T.U.J., Weidenbacher, P.A., Tang, S., Wu, W., Pak, J.E., and Kim, P.S. (2024). Efficient evolution of human antibodies from general protein language models. Nat Biotechnol 42, 275–283.

Homma, F., Huang, J., and van der Hoorn, R.A.L. (2023). AlphaFold-Multimer predicts cross-kingdom interactions at the plant-pathogen interface. Nat Commun 14, 6040.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., and Abecasis, G.R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44, 955–959.

Hu, J., Fan, J., Sun, Z., and Liu, S. (2020). NextPolish: a fast and efficient genome polishing tool for long-read assembly. Bioinformatics 36, 2253–2255.

Hu, J., Wang, Z., Sun, Z., Hu, B., Ayoola, A.O., Liang, F., Li, J., Sandoval, J.R., Cooper, D.N., Ye, K., Ruan, J., et al. (2023). An efficient error correction and accurate assembly tool for noisy long reads. bioRxiv, 2023, 2003: 531669.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., et al. (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res 35, W169–W175.

Huang, J., Lin, Q., Fei, H., He, Z., Xu, H., Li, Y., Qu, K., Han, P., Gao, Q., Li, B., et al. (2023). Discovery of deaminase functions by structure-based protein clustering. Cell 186, 3182–3195.e14.

Huang, M., Liu, X., Zhou, Y., Summers, R.M., and Zhang, Z. (2019). BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience 8.

Huang, N., and Li, H. (2023). compleasm: a faster and more accurate reimplementation of BUSCO. Bioinformatics 39.

Huynh-Thu, V.A., Irrthum, A., Wehenkel, L., and Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS One 5.

Jiang, L., Zheng, Z., Qi, T., Kemper, K.E., Wray, N.R., Visscher, P.M., and Yang, J. (2019). A resource-efficient tool for mixed model association analysis of large-scale data. Nat Genet 51, 1749–1755.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res 28, 27–30.

Kang, H.M., Sul, J.H., Service, S.K., Zaitlen, N.A., Kong, S., Freimer, N.B., Sabatti, C., and Eskin, E. (2010). Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42, 348–354.

Kang, H.M., Zaitlen, N.A., Wade, C.M., Kirby, A., Heckerman, D., Daly, M.J., and Eskin, E. (2008). Efficient control of population structure in model organism association mapping. Genetics 178, 1709–1723.

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S.O., and Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC BioInf 19, 189.

Keilwagen, J., Wenk, M., Erickson, J.L., Schattat, M.H., Grau, J., and Hartung, F. (2016). Using intron position conservation for homology-based gene prediction. Nucleic Acids Res 44, e89.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol 37, 907–915.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14, R36.

Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P.A. (2019). Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol 37, 540–546.

Koren, S., Rhie, A., Walenz, B.P., Dilthey, A.T., Bickhart, D.M., Kingan, S.B., Hiendleder, S., Williams, J.L., Smith, T.P.L., and Phillippy, A.M. (2018). De novo assembly of haplotype-resolved genomes with trio binning. Nat Biotechnol 36, 1174–1182.

Koren, S., Walenz, B.P., Berlin, K., Miller, J.R., Bergman, N.H., and Phillippy, A.M. (2017). Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. Genome Res 27, 722–736.

Korf, I. (2004). Gene finding in novel genomes. BMC BioInf 5, 59.

Krizanovic, K., Echchiki, A., Roux, J., and Sikic, M. (2018). Evaluation of tools for long read RNA-seq splice-aware alignment. Bioinformatics 34, 748–754.

Krueger, F., and Andrews, S.R. (2011). Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27, 1571–1572.

Lai, X., Behera, S., Liang, Z., Lu, Y., Deogun, J.S., and Schnable, J.C. (2017). STAG-CNS: an order-aware conserved noncoding sequences discovery tool for arbitrary numbers of species. Mol Plant 10, 990–999.

Langfelder, P., and Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. BMC BioInf 9, 559.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359.

Langmead, B., Wilks, C., Antonescu, V., and Charles, R. (2019). Scaling read aligners to hundreds of threads on general-purpose processors. Bioinformatics 35, 421–432.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100.

Li, H. (2021). New strategies to improve minimap2 alignment accuracy. Bioinformatics 37, 4572–4574.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754–1760.

Li, L., Stoeckert Jr., C.J., and Roos, D.S. (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13, 2178–2189.

Li, M., Liu, X., Bradbury, P., Yu, J., Zhang, Y.M., Todhunter, R.J., Buckler, E.S., and Zhang, Z. (2014). Enrichment of statistical power for genome-wide association studies. BMC Biol 12, 73.

Li, Y., Ge, X., Peng, F., Li, W., and Li, J.J. (2022). Exaggerated false positives by popular differential expression methods when analyzing human population samples. Genome Biol 23, 79.

Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30, 923–930.

Lipka, A.E., Tian, F., Wang, Q., Peiffer, J., Li, M., Bradbury, P.J., Gore, M.A., Buckler, E S., and Zhang, Z. (2012). GAPIT: genome association and prediction integrated tool. Bioinformatics 28, 2397–2399.

Lippert, C., Listgarten, J., Liu, Y., Kadie, C.M., Davidson, R.I., and Heckerman, D. (2011). FaST linear mixed models for genome-wide association studies. Nat Methods 8, 833–835.

Liu, Q., Wang, C., Jiao, X., Zhang, H., Song, L., Li, Y., Gao, C., and Wang, K. (2019). Hi-TOM: a platform for high-throughput tracking of mutations induced by CRISPR/Cas systems. Sci China Life Sci 62, 1–7.

Liu, X., Huang, M., Fan, B., Buckler, E.S., and Zhang, Z. (2016). Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12, e1005767.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic

Acids Res 33, 6494–6506.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15, 550.

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., He, G., Chen, Y., Pan, Q., Liu, Y., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. GigaScience 1, 18.

Luo, R., Sedlazeck, F.J., Lam, T.W., and Schatz, M.C. (2019). A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun 10, 998.

Luo, Y., Jiang, G., Yu, T., Liu, Y., Vo, L., Ding, H., Su, Y., Qian, W.W., Zhao, H., and Peng, J. (2021). ECNet is an evolutionary context-integrated deep learning framework for protein engineering. Nat Commun 12, 5743.

Ma, W., Qiu, Z., Song, J., Li, J., Cheng, Q., Zhai, J., and Ma, C. (2018). A deep convolutional neural network approach for predicting phenotypes from genotypes. Planta 248, 1307–1318.

Ma, X.K., Wang, M.R., Liu, C.X., Dong, R., Carmichael, G.G., Chen, L.L., and Yang, L. (2019). CIRCexplorer3: a clear pipeline for direct comparison of circular and linear RNA expression. Genomics Proteomics BioInf 17, 511–521.

Mao, Y. (2019). GenoDup Pipeline: a tool to detect genome duplication using the dS-based method. PeerJ 7, e6303.

Marcais, G., Delcher, A.L., Phillippy, A.M., Coston, R., Salzberg, S.L., and Zimin, A. (2018). MUMmer4: a fast and versatile genome alignment system. PLoS Comput Biol 14, e1005944.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., et al. (2010). The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res 20, 1297–1303.

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R.S., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol 17, 122.

Memczak, S., Jens, M., Elefsinioti, A., Torti, F., Krueger, J., Rybak, A., Maier, L., Mackowiak, S.D., Gregersen, L.H., Munschauer, M., et al. (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. Nature 495, 333–338.

Mendes, F.K., Vanderpool, D., Fulton, B., and Hahn, M.W. (2021). CAFE 5 models variation in evolutionary rates among gene families. Bioinformatics 36, 5516–5518.

Molder, F., Jablonski, K.P., Letcher, B., Hall, M.B., Tomkins-Tinch, C.H., Sochat, V., Forster, J., Lee, S., Twardziok, S.O., Kanitz, A., et al. (2021). Sustainable data analysis with Snakemake. F1000Res 10, 33.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R., et al. (2007). New developments in the InterPro database. Nucleic Acids Res 35, D224–D228.

Naito, Y., Hino, K., Bono, H., and Ui-Tei, K. (2015). CRISPRdirect: software for designing CRISPR/Cas guide RNA with reduced off-target sites. Bioinformatics 31, 1120–1123.

Nurk, S., Walenz, B.P., Rhie, A., Vollger, M.R., Logsdon, G.A., Grothe, R., Miga, K.H., Eichler, E.E., Phillippy, A.M., and Koren, S. (2020). HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. Genome Res 30, 1291–1305.

Ou, S., Chen, J., and Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res 46, e126.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R.A., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., et al. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol 20, 275.

Pandi, A., Adam, D., Zare, A., Trinh, V.T., Schaefer, S.L., Burt, M., Klabunde, B., Bobkova, E., Kushwaha, M., Foroughijabbari, Y., et al. (2023). Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. Nat Commun 14, 7197.

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14, 417–419.

Pérez-Enciso, M., Ramírez-Ayala, L.C., and Zingaretti, L.M. (2020). SeqBreed: a python tool to evaluate genomic prediction in complex scenarios. Genet Sel Evol 52, 7.

Pérez-Rodríguez, P., and de los Campos, G. (2022). Multitrait Bayesian shrinkage and variable selection models with the BGLR-R package. Genetics 222.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., and Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol 33, 290–295.

Poplin, R., Chang, P.C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., Newburger, D., Dijamco, J., Nguyen, N., Afshar, P.T., et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol 36, 983–987.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904–909.

Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics 155, 945–959.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes de novo assembler. CP BioInf 70, e102.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res 33, W116–W120.

Rhie, A., Walenz, B.P., Koren, S., and Phillippy, A.M. (2020). Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol 21, 245.

Riesselman, A.J., Ingraham, J.B., and Marks, D.S. (2018). Deep generative models of genetic variation capture the effects of mutations. Nat Methods 15, 816–822.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C.L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci USA 118, e2016239118.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Salikhov, K., Sacomoto, G., and Kucherov, G. (2014). Using cascading Bloom filters to improve the memory usage for de Brujin graphs. Algorithms Mol Biol 9, 2.

Scheet, P., and Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78, 629–644.

Schmitt, L.T., Paszkowski-Rogacz, M., Jug, F., and Buchholz, F. (2022). Prediction of designer-recombinases for DNA editing with generative deep learning. Nat Commun 13, 7966.

Segura, V., Vilhjálmsson, B.J., Platt, A., Korte, A., Seren, Ü., Long, Q., and Nordborg, M. (2012). An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44, 825–830.

Senior, A.W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A.W.R., Bridgland, A., et al. (2020). Improved protein structure prediction using potentials from deep learning. Nature 577, 706–710.

Simao, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., and Zdobnov, E.M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212.

Slater, G.S.C., and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC BioInf 6, 31.

Smolka, M., Paulin, L.F., Grochowski, C.M., Horner, D.W., Mahmoud, M., Behera, S., Kalef-Ezra, E., Gandhi, M., Hong, K., Pehlivan, D., et al. (2024). Detection of mosaic and population-level structural variants with Sniffles2. Nat Biotechnol doi: 10.1038/s41587-023-02024-y.

Solovyev, V., Kosarev, P., Seledsov, I., and Vorobyev, D. (2006). Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biol 7, S10–11.

Song, B., Buckler, E.S., Wang, H., Wu, Y., Rees, E., Kellogg, E.A., Gates, D.J., Khaipho-Burch, M., Bradbury, P.J., Ross-Ibarra, J., et al. (2021). Conserved noncoding sequences provide insights into regulatory sequence and loss of gene expression in maize. Genome Res 31, 1245–1257.

Song, B., Marco-Sola, S., Moreto, M., Johnson, L., Buckler, E.S., and Stitzer, M.C (2022). AnchorWave: sensitive alignment of genomes with high sequence diversity, extensive structural polymorphism, and whole-genome duplication. Proc Natl Acad Sci USA 119, e2113075119.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006). AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res 34, W435–W439.

Stein, L.D. (2013). Using GBrowse 2.0 to visualize and share next-generation sequence data. Brief BioInf 14, 162–171.

Strodthoff, N., Wagner, P., Wenzel, M., and Samek, W. (2020). UDSMProt: universal deep sequence models for protein classification. Bioinformatics 36, 2401–2409.

Stuart, T., Srivastava, A., Madad, S., Lareau, C.A., and Satija, R. (2021). Single-cell chromatin state analysis with Signac. Nat Methods 18, 1333–1341.

Sun, P., Jiao, B., Yang, Y., Shan, L., Li, T., Li, X., Xi, Z., Wang, X., and Liu, J. (2022). WGDI: a user-friendly toolkit for evolutionary analyses of whole-genome duplications and ancestral karyotypes. Mol Plant 15, 1841–1851.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science 320, 486–488.

Tang, Y., Liu, X., Wang, J., Li, M., Wang, Q., Tian, F., Su, Z., Pan, Y., Liu, D., Lipka, A. E., et al. (2016). GAPIT version 2: an enhanced integrated tool for genomic association and prediction. Plant Genome 9.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics 31, 2032–2034.

Thiel, T., Kota, R., Grosse, I., Stein, N., and Graner, A. (2004). SNP2CAPS: a SNP and INDEL analysis tool for CAPS marker development. Nucleic Acids Res 32, 5e–5.

Thorvaldsdottir, H., Robinson, J.T., and Mesirov, J.P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief BioInf 14, 178–192.

Tian, T., Liu, Y., Yan, H., You, Q., Yi, X., Du, Z., Xu, W., and Su, Z. (2017). agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res 45, W122–W129.

Torres, S.V., Leung, P.J.Y., Venkatesh, P., Lutz, I.D., Hink, F., Huynh, H.H., Becker, J., Yeh, A.H.W., Juergens, D., Bennett, N.R., et al. (2024). De novo design of high-affinity binders of bioactive helical peptides. Nature 626, 435–442.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J., and Pachter, L. (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28, 511–515.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B.C., Remm, M., and Rozen, S.G. (2012). Primer3—new capabilities and interfaces. Nucleic Acids Res 40, e115.

Usadel, B., Nagel, A., Steinhauser, D., Gibon, Y., Bläsing, O.E., Redestig, H., Sreenivasulu, N., Krall, L., Hannah, M.A., Poree, F., et al. (2006). PageMan: an interactive ontology tool to generate, display, and annotate overview graphs for profiling experiments. BMC BioInf 7, 535.

Usadel, B., Poree, F., Nagel, A., Lohse, M., Czedik-eysenberg, A., and Stitt, M. (2009). A guide to using MapMan to visualize and compare Omics data in plants: a case study in the crop species, Maize. Plant Cell Environ 32, 1211–1229.

Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 27, 737–746.

Vasimuddin, M., Misra, S., Li, H., and Aluru, S. (2019). Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems, pp. 314–324.

Voichek, Y., and Weigel, D. (2020). Identifying genetic variants underlying phenotypic variation in plants without complete genomes. Nat Genet 52, 534–540.

Voorrips, R.E., and Maliepaard, C.A. (2012). The simulation of meiosis in diploid and tetraploid organisms using various genetic models. BMC BioInf 13, 248.

Walker, B.J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S.K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS ONE 9, e112963.

Wang, J., Lisanza, S., Juergens, D., Tischer, D., Watson, J.L., Castro, K.M., Ragotte, R., Saragovi, A., Milles, L.F., Baek, M., et al. (2022). Scaffolding protein functional sites using deep learning. Science 377, 387–394.

Wang, J., and Zhang, Z. (2021). GAPIT version 3: boosting power and accuracy for genomic association and prediction. Genomics Proteomics BioInf 19, 629–640.

Wang, K., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S., and Li, H. (2023). DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. Mol Plant 16, 279–293.

Wang, Y., Tang, H., DeBarry, J.D., Tan, X., Li, J., Wang, X., Lee, T., Jin, H., Marler, B., Guo, H., et al. (2012). MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40, e49.

Wang, Y., You, Z.H., Yang, S., Li, X., Jiang, T.H., and Zhou, X. (2019). A high efficient biological language model for predicting protein-protein interactions. Cells 8, 122.

Wei, X., Qiu, J., Yong, K., Fan, J., Zhang, Q., Hua, H., Liu, J., Wang, Q., Olsen, K.M., Han, B., et al. (2021). A quantitative genomics map of rice provides genetic insights and guides breeding. Nat Genet 53, 243–253.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018.

Xi, Y., and Li, W. (2009). BSMAP: whole genome bisulfite sequence MAPping program. BMC BioInf 10, 232.

Xie, X., Ma, X., Zhu, Q., Zeng, D., Li, G., and Liu, Y.G. (2017). CRISPR-GE: a convenient software toolkit for CRISPR-based genome editing. Mol Plant 10, 1246–1249.

Xu, Z., and Wang, H. (2007). LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res 35, W265–W268.

Yang, A., Jude, K.M., Lai, B., Minot, M., Kocyla, A.M., Glassman, C.R., Nishimiya, D., Kim, Y.S., Reddy, S.T., Khan, A.A., et al. (2023). Deploying synthetic coevolution and machine learning to engineer protein-protein interactions. Science 381, eadh1720.

Yang, J., Anishchenko, I., Park, H., Peng, Z., Ovchinnikov, S., and Baker, D. (2020). Improved protein structure prediction using predicted interresidue orientations. Proc Natl Acad Sci USA 117, 1496–1503.

Yang, Y., Li, Y., Chen, Q., Sun, Y., and Lu, Z. (2019). WGDdetector: a pipeline for detecting whole genome duplication events using the genome or transcriptome annotations. BMC BioInf 20, 75.

Yu, J., Pressoir, G., Briggs, W.H., Vroh Bi, I., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Holland, J.B., et al. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38, 203–208.

Zhang, J., Chen, S., Yang, J., and Zhao, F. (2020). Accurate quantification of circular RNAs identifies extensive circular isoform switching events. Nat Commun 11, 90.

Zhang, X., Zhang, S., Zhao, Q., Ming, R., and Tang, H. (2019). Assembly of allele-aware, chromosomal-scale autopolyploid genomes based on Hi-C data. Nat Plants 5, 833–845.

Zhang, X.O., Dong, R., Zhang, Y., Zhang, J.L., Luo, Z., Zhang, J., Chen, L.L., and Yang, L. (2016). Diverse alternative back-splicing and alternative splicing landscape of circular RNAs. Genome Res 26, 1277–1287.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based analysis of ChIP-Seq (MACS). Genome Biol 9, R137.

Zhang, Z., Ersoz, E., Lai, C.Q., Todhunter, R.J., Tiwari, H.K., Gore, M.A., Bradbury, P.J., Yu, J., Arnett, D.K., Ordovas, J.M., et al. (2010). Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42, 355–360.

Zhao, N., and Boyle, A.P. (2021). F-Seq2: improving the feature density based peak caller with dynamic statistics. NAR Genomics BioInf 3, lqab012.

Zhou, Q., Lim, J.Q., Sung, W.K., and Li, G. (2019). An integrated package for bisulfite DNA methylation data analysis with Indel-sensitive mapping. BMC BioInf 20, 47.

Zhou, X., Chen, G., Ye, J., Wang, E., Zhang, J., Mao, C., Li, Z., Hao, J., Huang, X., Tang, J., et al. (2023). ProRefiner: an entropy-based refining strategy for inverse protein folding with global graph attention. Nat Commun 14, 7434.

Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44, 821–824.

Zwaenepoel, A., and Van de Peer, Y. (2019). WGD—simple command line tools for the analysis of ancient whole-genome duplications. Bioinformatics 35, 2153–2155.

https://doi.org/10.1007/s11427-024-2578-6