Genomics and transcriptomics advance in plant sciences

Boas Pucker[1*] (0000-0002-3321-7471) and Hanna Marie Schilbert[1] (0000-0003-0474-7753)

1 Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany

Abstract

Recent progress in sequencing technologies facilitates plant science experiments through the availability of genome and transcriptome sequences. Genome assemblies provide details about genes, transposable elements, and the general genome structure. The availability of a reference genome sequence for a species enables and supports numerous wet lab analyses and comprehensive bioinformatic investigations e.g. genome-wide investigations of gene families. After generating a genome sequence, gene prediction and the generation of functional annotations are the major challenges. Although these methods were improved substantially over the last years, incorporation of external hints like RNA-Seq reads is beneficial. Once a high-quality sequence and annotation is available for a species, diversity between accessions can be assessed by re-sequencing. This helps in revealing single nucleotide variants, insertions and deletions, and larger structural variants like inversions and transpositions. Identification of these variants requires sophisticated bioinformatic tools and many of them were developed during past years. Sequence variants can be harnessed for the genetic mapping of traits. Several mapping-by-sequencing approaches were developed to find underlying genes for relevant traits in crops. These genomic approaches are complemented by various transcriptomic methods dominated by a very popular RNA-Seq technology. Transcript abundance is measured via sequencing of the corresponding cDNA molecules. RNA-Seq reads can be subjected to transcriptome assembly or gene expression analysis, e.g. for the identification of transcripts abundance between different tissues, conditions, or genotypes.

Key words: Bioinformatics, Computational biology, Sequencing, Genome assembly, Gene prediction, Read mapping, Variant calling, Single nucleotide variant (SNV), Insertion/deletion (InDel), Genotyping-by-sequencing (GBS), Mapping-by-sequencing (MBS), RNA-Seq, Transcriptome assembly

Published version of this manuscript:

Pucker B & Schilbert H. Genomics and Transcriptomics Advance in Plant Sciences. Molecular Approaches in Plant Biology and Environmental Challenges. Springer. 2019. ISBN 978-981-15-0690-1. doi:10.1007/978-981-15-0690-1.

Introduction

The genome of an organism determines its phenotype by setting the range of variability for numerous traits. Environmental factors shape the phenotype within this predetermined range. Knowledge about the genome and genes of a species facilitates various biological research projects. Research on *Arabidopsis thaliana* (*A. thaliana*) Columbia-0 was boosted by the availability of the first plant genome sequence [1]. The transcriptome of an organism reveals which parts of the genome are 'active' at a certain point in time, under specific conditions, and in a defined cell type. Since the nucleic acid types DNA and RNA have very similar biochemical properties, the investigation of genome and transcriptome can be performed by similar methods. Both omics layers, genomics and transcriptomics, are easily accessible by analytic methods, because general biochemical properties of these nucleic acids are independent from the actual sequence. The intention of this chapter is 1) to describe genomics and transcriptomics workflows which are commonly used in plant research, and 2) to list frequently deployed bioinformatic tools for the analysis steps (Fig. 1).

Sequencing technologies

Existing sequencing technologies can be grouped into different generations based on their key properties. However, there is disagreement in the literature about this classification system and the assignment of technologies to different generations [2–9]. Here, we distinguish between three generations: I) Sanger chain termination sequencing and Maxam Gilbert sequencing as first generation sequencing technologies, II) Roche/454 pyrosequencing, IonTorrent, Solexa/Illumina, and Beijing Genomics Institute (BGI) sequencing as second generation sequencing technologies, and III) Single molecule real time sequencing (Pacific Bioscience, PacBio) and nanopore sequencing (Oxford Nanopore Technologies, ONT) as third generation sequencing technologies. Technical details of these sequencing technologies were reviewed elsewhere [2,4,7,8,10–12].

Since the invention of chain termination sequencing [13,14], substantial technological advances paved the way for cost reductions. Therefore, broad application of high throughput sequencing [2] and more recently long read sequencing technologies [15] became possible. Sanger sequencing generates a single read per sample, while other technologies produce large amounts of reads per sample and are hence crucial for many genome sequencing projects. Length of reads produced from Roche 454 pyrosequencing and IonTorrent is comparable to Sanger sequencing, but have reduced accuracy. Nevertheless, For years, Illumina has been dominating the market for high throughput sequencing with substantially short reads due to high accuracy and low costs of sequencing technology. The BGI became a serious competitor during past years and is now offering the generation of similar sequencing data-sets based on its own technologies. While Illumina sequencing platforms are distributed all around the globe, BGI sequencing technology is exclusively available in China.

Paired-end sequencing provides the opportunity to analyze two ends of the same molecule. Overlapping reads; e.g. 2x300nt, can be merged, thus leading to a total length of up to 500 nt. Sophisticated approaches like TrueSeq synthetic Long-Reads [16] were developed to maximize the read length of second generation technologies up to several thousand nucleotides. Mate pair reads provide information about the distance of both reads in addition to the mere sequences of both reads. In mate pair sequencing technique, long DNA fragments are modified at their ends, circularized, and fragmented. Fragments with marks are enriched and finally sequenced as paired-end libraries. The size of the initial fragments determines the distance of the two generated reads and can thus be considered valuable linkage information during genome assembly processes.

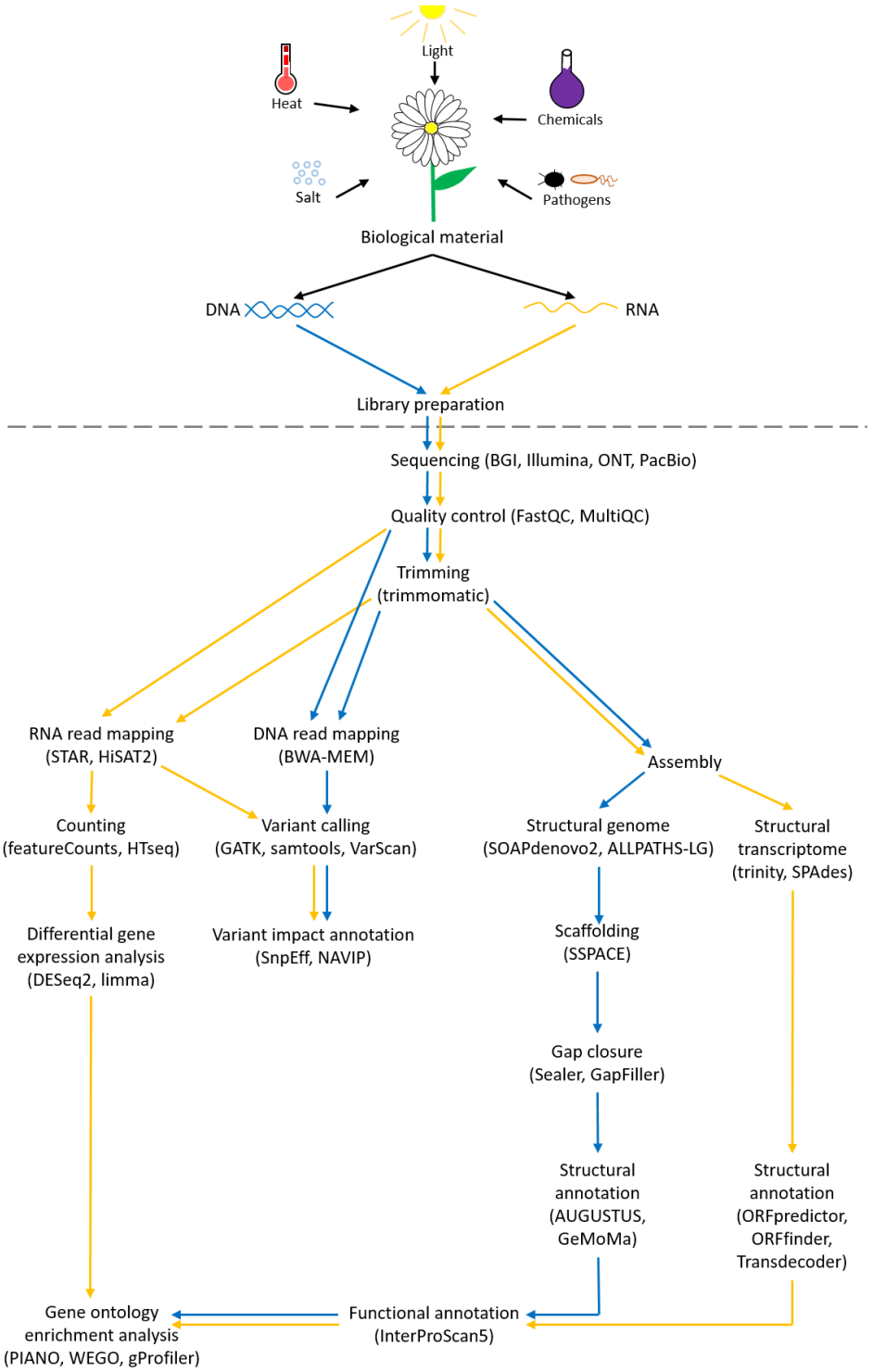

Fig.1: Selected genomics and transcriptomics workflows in plant sciences. These workflows are deployed in many studies in plant research and the listed tools can be applied to perform the displayed steps. Several alternative and additional tools are listed within this chapter.

However, length of reads generated from mate-pair sequencing is inferior to those generated by Oxford Nanopore Technologies (ONT) and Pacific Biosciences. From ONT, the longest sequenced DNA molecule has been reported to be over 2 Mbp till date [17] and the longest single reads is close to 1 Mbp [18]. Dropping sequencing costs and the rise of long read technologies enabled sequencing projects for numerous plant species [19–21]. Nevertheless, short reads are still valuable in projects; e.g. RNA-Seq or re-sequencing projects, where a high number of tags is more important than the read length.

In addition to generating extremely long reads at low costs, ONT also provides the first portable sequencers, namely MinION and Flongel, that can be deployed in field applications [22,23]. Sequencing in the field opens up opportunities, to monitor pathogens in the field accurately [24] and to assess the biodiversity [23]. Real time base calling and the start of downstream analysis before completion of a sequencing run are beneficial when decisions are time critical [25]. Moreover, it also allows researchers to stop the sequencing process once sufficient data is generated and to commit the remaining sequencing capacity to other projects [26].

Genomics

Genome assembly

Quality control and preprocessing. Quality checks via FastQC [27] or MultiQC [28] are usually the first steps to assess the quality of sequencing data. Next, reads need to be preprocessed prior to a *de novo* assembly, while this is not necessary for other applications like read mapping. Low quality sequences and remaining adapter fragments are removed during the trimming process, e.g. by trimmomatic [29]. Removal of adapter sequences is especially important for *de novo* genome assemblies, because these sequences can occur in independent reads and cause the miss-join of random sequences into contigs.

Assembly concept. A read can only represent a fraction of a complete genome sequence. Hence, intense manual work or the application of sophisticated bioinformatic tools is necessary to reconstruct complete genome sequences based on sequence reads [30–32]. Initial sequencing projects involved the cloning of genomic fragments into vectors like bacterial artificial chromosomes (BACs) prior to sequencing. Genome sequences were generated by sequencing several BACs consecutively and combining the BAC sequences almost manually.

Second generation genome assemblies. Especially, the rise of high throughput sequencing methods caused a shift from manually curated BAC-based high continuity genome sequences towards whole genome shotgun draft assemblies. Dedicated assemblers were developed to harness the full potential of the available data types, for example combinations of paired-end and mate-pair data. SOAPdenovo2 [33], ALLPATHS-LG [34], Platanus [35], and the proprietary CLC assembler [36] are examples for tools which were successfully deployed for the assembly of plant genomes, but there are also many alternatives (Table 1). Modification of parameters, especially *k*-mer sizes, should be optimized empirically [37–40]. In addition, the best combination of data from multiple sequencing libraries and sequencing technologies needs to be identified. After the generation of contigs in the assembly process, the information of mate pair and paired-end data-sets can be used to connect contigs to scaffolds without knowing the sequence enclosed between contigs of a scaffold. While some assemblers provide this functionality, dedicated tools like SSPACE [41] are available. Next, gaps between contigs within a scaffold can be partially closed, e.g. via GapFiller [42] or Sealer [43]. The reduced sequencing costs allowed the assembly of plant genome sequences by single groups [44], but most genome sequences were highly fragmented. More recently, the proprietary NRGene assembler (DeNovoMAGIC^{TM}) and the competing open source alternative TRITEX [45] are promising substantially improved assemblies.

Table 1: Assembler for second generation sequencing data. This table is an **i**ncomplete list of tools that can be applied for the *de novo* plant genome assembly based on second generation sequencing data.

Name	Availability	Link	Reference
CLC	Licence required	https://www.qiagenbioinformatics.com/products/clc-main-workbench	[36]
SOAPdenovo2	Binary available	https://github.com/aquaskyline/SOAPdenovo2	[33]
Velvet	Installation required	https://github.com/dzerbino/velvet	[46]
ALLPATHS-LG	Installation required	http://software.broadinstitute.org/allpaths-lg/blog/?page_id=12	[34]
Ray	Installation required	http://denovoassembler.sourceforge.net	[47]
Newbler	Installation required	http://sequencing.roche.com	[11]
MaSuRCA	Installation required	https://github.com/alekseyzimin/masurca	[48]
SGA	Installation required	https://github.com/jts/sga	[49]
Platanus	Installation required	http://platanus.bio.titech.ac.jp	[35]

Third generation genome assemblies. The assembly situation changed again when long reads became available, thus enabling the generation of high continuity genome assemblies for numerous plant species with moderate effort [50–53]. The technological boost on the sequencing side caused an explosion in the development of novel assemblers and read correction tools which can handle noisy long reads efficiently (Table 2). FALCON [54], Canu [55], Flye [56], Miniasm [57], and wtdbg2 [58] are examples for frequently applied assemblers. Depending on the sequencing coverage and repeat content, the computational costs of assemblies can be high. Several hundred CPU hours, some hundred GB of RAM, and several TB of disc space are often required to assemble plant genomes. Assembled contigs can be joined into scaffolds based on additional information like genetic linkage [51,59], optical mapping information, e.g. from Bionano Genomics and OptGen [60–62], and Hi-C [60,63,64]. Genetic linkage can rely on molecular markers measured in the lab [51] or on sequencing of multiple individual plants of a segregating population by a high throughput method [59]. Optical mapping is a size estimation of large DNA fragments which are generated by enzymatic restriction digest and cut site specific coloring with fluorescent dyes. Hi-C measures the 3D distances of genomic loci and assumes that neighboring sequences are also likely to be co-located in 2D.

Due to the high error rate in long reads, raw assemblies require several polishing steps. Firstly, long reads are aligned for correction, e.g. via BLASR [65] and minimap2 [66]. Arrow [54] can be applied to polish assemblies based on PacBio reads, while nanopolish [67] is the best choice for ONT reads. Secondly, highly accurate short reads are mapped to the assembly to further correct the sequence in single copy regions. Paired-end or mate pair reads provide higher specificity during the mapping compared to single end reads. BWA-MEM [68] is a suitable read mapping tool and Pilon [69] can be used for the detection and correction of assembly errors. Iterative rounds of correction are possible. There is still an ongoing debate about the optimal number of polishing rounds that should be performed [55,70]. Since the most frequent error types are insertions/deletions, open reading frames are often affected by apparent frameshifts and premature stop codons. Therefore, the contribution of polishing approaches can be benchmarked based on an increase/decrease of frameshifts and premature stop codons in protein encoding genes. The optimal number of correction rounds can be determined by minimizing the number of these variants.

Table 2: Third generation assembler. This table is an incomplete list of tools that can be applied for the *de novo* plant genome assembly based on third generation sequencing data.

Name	Availability	Link	Reference
FALCON	SMRT Link	https://www.pacb.com/training/smrt-link-overview	[54]
Canu	Installation required	https://github.com/marbl/canu	[55]
Flye	Installation required	https://github.com/fenderglass/Flye	[56]
Miniasm	Installation required	https://github.com/lh3/miniasm	[57]
wtdbg2	Installation required	https://github.com/ruanjue/wtdbg2	[58]

Assembly validation. After combining reads into contigs, the correctness of these connections needs to be assessed. This assembly validation can be performed by mapping all reads back to the generated sequence, e.g. via BWA-MEM [68], and analyzing the distances of paired reads in this mapping, e.g. via REAPR [71]. Alternative approaches like implemented in KAT [72] inspect the assembly based on included *k*-mers. Most genome sequencing projects involve the generation of multiple assemblies with different tools and parameter settings. Selection of the best assembly can be challenging and criteria depend on the proposed research questions. The largest reasonable assembly, the assembly with the highest continuity, or the assembly resolving the highest number of genes might be of interest. Benchmarking Universal Single‐Copy Orthologs (BUSCO) [73] is a frequently applied method to assess the assembly completeness and correctness. The underlying assumption is that all benchmarking genes should appear exactly once in the assembly. Different benchmarking sets exist for different taxonomic groups [74]. Due to a large phylogenetic distance to other sequenced species, this might not be perfectly accurate for the species of interest. However, the detection of single copy and complete genes is a good indicator for a high quality assembly. High numbers of duplicated BUSCOs can indicate separated haplophases. Recently, DOGMA [75] was released as an alternative tool for the analysis of sequence set completeness which also comes with an online version (https://domainworld-services.uni-muenster.de/dogma).

Gene prediction

After generation and polishing of an assembly, the prediction of genes is often the next step. Besides protein encoding genes, there are also various RNA genes, transposable element genes, and numerous repeats which should be annotated as part of a genome project. In general, predictions are distinguished into I) intrinsic approaches, which rely only on sequence properties, and II) extrinsic approaches, which harness sequence similarity to previously annotated sequences to transfer annotation. However, frequently applied tools are designed to harness the power of both approaches (Table 3). AUGUSTUS [76,77] and GeneMark derivatives [78–81] can predict genes *ab initio* without any external information. BUSCO can be applied to generate parameter files for this gene prediction process by assessing the gene structure of BUSCO genes [82]. In contrast to these *ab initio* approaches, GeMoMa [83,84] combines external hints to construct a gene annotation based on sequence alignments. The exon intron structure of plant genes is posing a challenge to the gene prediction process, because tools need to account for interruptions of an open reading frame by on average four to five introns per gene [85]. Intron borders are often detected based on their conserved sequences: GT at the 5’ end and AG at the 3’ end. However, an average of at least 5% of all plant genes contains non‐canonical splice sites, i.e. deviations from the GT-AG combination [85,86]. Most gene prediction tools exclude non‐canonical splice sites at least in the *ab initio* mode, because the number of possible gene models increases substantially when permitting many more possible intron positions. Therefore, external hints for intron positions are crucial to achieve an accurate prediction. If the identification of all isoforms of a gene is of interest, the accurate annotation of all exon intron borders is especially important. Expressed sequence tags (ESTs), contigs of a transcriptome assembly, or unassembled RNA-Seq reads can be aligned to the genomic sequence to generate hints. These sequences should originate from a broad range of different samples, e.g. collected under different environmental conditions, from different tissues, and different developmental stages. The accurate alignment of transcript sequences to an assembly requires dedicated tools to

account for introns. While BLAT [87] can align long sequences, STAR [88,89] is well suited for the split alignment of RNA-Seq reads. Dedicated tools like exonerate [90] allow the alignment of previously annotated peptide sequences from other species. Resulting alignments can be converted into gene prediction hints. Annotation pipelines like MAKER2 [91], BRAKER1 [92], and Gnomon [93] can integrate the information from different hint sources with *ab initio* prediction. While the prediction of protein encoding parts of a gene works relatively well, the annotation of untranslated regions (UTRs) and other non-coding sequences is still associated with a higher insecurity [86,94,95]. Quality of the gene prediction process is in general not keeping pace with the rapid improvement of sequencing capacities and the frequent generation of highly contiguous assemblies [96].

Technological progress allows the systematic investigation of non-protein encoding genes; e.g. through RNA-Seq experiments committed to the analysis of short RNAs. INFERNAL [97] and tRNAscan-SE2 [98] are tools for the prediction of pure RNA genes.

Masking of repeats, e.g. via RepeatMasker [99], is frequently performed prior to the prediction of protein encoding genes, but this can actually have almost no or even detrimental effects on the prediction accuracy of certain gene families [100]. Although transposable elements and other repeats account for the major proportion of many plant genomes [101,102], the annotation of repeats is often performed poorly or omitted completely [103–105]. There is a plethora of annotation tools like RepeatScout [106] and RepeatMasker [99]. Bioinformatic pipelines were developed to account for weaknesses of single tools and to combine the strengths of many individual tools [107–109]. One major issue with the TE and repeat annotation is the lack of a universal benchmarking study which could hint to the best tool for certain purposes [105,110]. While the annotation of protein encoding genes can be checked for completeness based on BUSCO [73] and DOGMA [75,111], there is no such benchmarking data-set available for TEs.

Table 3: Plant gene prediction tools. This table is an **i**ncomplete list of tools that can be applied for gene prediction on plant genome assemblies.

Name	Availability	Link	Reference
AUGUSTUS	Installation required	https://github.com/Gaius-Augustus/Augustus	[76]
BRAKER1	Installation required	https://github.com/Gaius-Augustus/BRAKER	[92]
GeneMark	Installation required	http://exon.gatech.edu/GeneMark/license_download.cgi	[79–81]
GeMoMa	Jar file	http://www.jstacs.de/index.php/GeMoMa	[83,84]
Gnomon	Installation required	ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools++/CURRENT	[93]
MAKER2	Registration required	https://www.yandell-lab.org/software/maker.html	[91]
SNAP	Installation required	https://github.com/KorfLab/SNAP	[112]

Application examples. Sequencing the genome of a plant species can provide insights into specific adaptations to local environmental conditions. *Crucihimalaya himalaica* is distributed at high altitudes at the Himalaya and the genome sequence reveals a reduced number of pathogen response genes as well as an increased number of DNA repair genes as response to a reduced amount of pathogens and an increased UV exposure, respectively [111].

Re-sequencing and variant calling

Once a suitable reference genome sequence is available, re-sequencing projects can by-pass the laborious and expensive assembly step. Reads can be mapped to a reference sequence to identify differences between individuals of the same species or even between closely related species. Since the re-sequencing dataset does not need to provide sufficient data for a *de novo* assembly, the costs for re-sequencing are low compared to the initial genome project. Re-sequencing of over 1,135 *A. thaliana* accessions revealed insights into the genomic diversity of

this species [113]. Since accessions are adapted to local environmental conditions, this project can reveal insights into adaptation mechanisms. Sequencing data also advances the understanding of population structures, genomic diversity between accessions, and genome evolution.

BWA-MEM [68] and bowtie2 [114] are frequently applied tools for the mapping of reads to a reference sequence (Table 4). The removal of PCR duplicates is necessary to avoid introducing a bias into following coverage analyses or variant callings. PCR duplicates are reads originating from a DNA fragment, which was amplified by PCR during the sequencing library preparation step. Functions like MarkDuplicates of Picard tools [115] allow the identification and removal of reads or read pairs originating from identical PCR products. This removal can be based on identical read sequences or identical positions in the mapping to a reference sequence. The detection of copy number variations depends on the equal representation of all genomic parts in the reads. PCR duplicates could cause the identification of false positive duplications by producing a high numbers of identical reads which could display an apparent variant caused by a PCR error in an early amplification step. The identification of sequence variants is sensitive to PCR duplicates, because a certain number of reads displaying a variant is frequently used as filter criteria to remove false positive variant calls.

Table 4: Read mapping tools. This table is an **i**ncomplete list of tools which can be applied to map reads from second generation sequencing technologies against a reference sequence. While some tools are suitable for the continuous alignment of DNA reads, others can generate split alignments for RNA-Seq reads.

Name	Availability	Link	DNA/RNA	Reference
BWA-MEM	Installation required	https://github.com/lh3/bwa	DNA	[68]
Bowtie 2	Installation required	https://github.com/BenLangmead/bowtie2	DNA	[114]
GEM 3	Installation required	https://github.com/smarco/gem3-mapper	DNA	[116]
bbmap	Jar file available	https://sourceforge.net/projects/bbmap	DNA	[117]
Novoalign	Trial available	http://www.novocraft.com/products/novoalign	DNA	[118]
NextGenMap	Installation required	https://github.com/Cibiv/NextGenMap	DNA	[119]
MAQ	Installation required	http://maq.sourceforge.net/maq-man.shtml	DNA	[120]
RMAP	Installation required	https://github.com/smithlabcode/rmap	DNA	[121]
MOSAIK	Installation required	https://github.com/wanpinglee/MOSAIK	DNA	[122]
segemehl	Installation required	https://www.bioinf.uni-leipzig.de/Software/segemehl	RNA	[123]
STAR	Installation required	https://github.com/alexdobin/STAR	RNA	[88]
HISAT2	Binary available	https://ccb.jhu.edu/software/hisat2/manual.shtml	RNA	[124]

There are numerous tools for the detection of genomic differences based on a short read mapping (Table 5). Genome Analysis Tool Kit (GATK) [125,126], samtools/bcftools [127], and VarDict [128] can detect single nucleotide variations (SNVs) and small insertions/deletions (InDels). The rise of long read sequencing technologies added substantially to the sensitivity of the insertion/deletion detection. Moreover, it allows the identification of large scale structural rearrangements. GraphMap [129], marginAlign [130], and PoreSeq [131] can align long reads to a reference sequence to call variants. Other tools like SVIM [132] rely on alignments generated by dedicated

long read aligners like minimap2 [66] or BLASR [65]. Identified variants can be subjected to downstream filtering; e.g. based on the number of supporting and contradicting reads.

Table 5: Variant callers. This table is an **i**ncomplete list of tools which can be applied to identify sequence variants based on reads mapped against a reference sequence. While some tools are restricted to the identification of small variants, other can detect large structural variants.

Name	Availability	Link	Variants	Reference
DeepVariant	Installation required	https://github.com/google/deepvariant	Small	[133]
GATK	Jar file	https://software.broadinstitute.org/gatk/download	Small	[125,126]
SNVer	Installation required	http://snver.sourceforge.net	Small	[134]
SAMtools	Jar file	http://samtools.sourceforge.net	Small	[127]
VarDict	Installation required	https://github.com/AstraZeneca-NGS/VarDict	Small	[128]
VarScan 2	Jar file	http://varscan.sourceforge.net	Small	[135]
LoFreq	Binary available	https://csb5.github.io/lofreq/installation	Small	[136]
Platypus	Installation required	https://github.com/andyrimmer/Platypus	Small	[137]
SOAPsnp	Installation required	https://sourceforge.net/projects/soapsnp	Small	[138]
Atlas-SNP2	Installation required	https://sourceforge.net/projects/atlas2	Small	[139]
FreeBayes	Installation required	https://github.com/ekg/freebayes	Small	[140]
SVIM	Installation required	https://github.com/eldariont/svim	Large	[132]
marginAlign	Installation required	https://github.com/benedictpaten/marginAlign	Large	[130]
GraphMap	Installation required	https://github.com/isovic/graphmap	Large	[129]
PoreSeq	Installation required	https://github.com/tszalay/poreseq	Large	[131]

Once the variants are identified, it is possible to assign functional annotations. Established tools for this purpose are SnpEff [141] and ANNOVAR [142]. Based on the structural annotation of the reference sequence, SnpEff and ANNOVAR assign functional implications like “premature stop codon” or “frameshift” to single variants. Since these tools are predicting the effect for a single variant at a time, NAVIP [143] was developed for the integrated annotation of all variants within one coding sequence. NAVIP accounts for combined effects of neighboring variants, e.g. two short InDels which are both causing a frameshift on their own, but result in a few substituted amino acids when considered together.

Mapping by sequencing

Forward genetics. Forward genetics describes the genetic screening of mutants which have been isolated based on an outstanding phenotype [144]. Crossing a mutant with a wild type plant and selfing of the F1 offspring leads to a segregating F2 population. A large segregating population forms the basis for a forward genetics screen. Such a population contains members with the wild-type and mutant phenotypes, respectively. Except for the causal locus, the genotypes of this population should display a random distribution of alleles. Since this population is used for genetic mapping, it is called a mapping population. Genetic markers located near the causal mutation will co-segregate with this mutation. As a result of this linkage between the causal locus and flanking markers, one allele of the flanking markers should be over-represented in the mutant plants. Due to a gradually decreasing linkage, the frequency of the coupled marker allele should drop when moving away from the causal locus. Therefore, the allele frequency can be used to pinpoint loci of interest. Originally, the identification of the location of the causal mutation in the genome of a mutant has been a long-lasting procedure requiring a high number of genetic markers. Once a target region has been identified, this region was screened for candidate genes. In order to validate the link between the assumed candidate gene and the expected phenotype, complementation

experiments were frequently conducted. In following studies, the molecular function of the mutated gene was often elucidated.

Next generation forward genetics. Technological advances in next generation sequencing enable the use of small sequence variants as genetic markers. Since these small sequence variants occur in large numbers, the resolution of the resulting genetic map is extremely high. Allele frequencies at all sequence variants are calculated for identification of genomic regions associated with the phenotype of interest [145]. First approaches used bulk segregant analysis (BSA), where DNA from the mapping population is pooled based on the phenotypes of individuals and then sequenced, i.e. one pool comprises the wild type allele of a certain locus and the other pool the mutant allele of the respective locus. Next, reads are mapped against a reference genome sequence to detect sequence variants. In the next step, allele frequencies for all small sequence variants are calculated. High allele frequencies can indicate linkage with the causal locus. This approach is also known as mapping-by-sequencing (MBS) and allows the fast and simple identification of causal mutations through allele frequency deviations [144].

Mutagenesis. Natural variation can provide mutants, but it is also possible to generate mutant plants via mutagenesis. DNA damaging agents deployed in these mutagenesis experiments can be classified as physical mutagens (e.g. gamma radiation and fast neutron bombardment) or chemical mutagens (e.g. ethyl methanesulfonate, diepoxybutane, sodium azide) [146]. In order to achieve maximal genetic variation with a minimum decrease in viability, mutagenic dosage and specific properties of the mutagen need to be considered [146]. High mutagenic dosages likely result in a high number of mutations in the individual genome, thus the high diversity around a causal mutation might impede the identification [144]. If a mutagen introduces large genomic rearrangements (e.g. deletions or translocation of large regions), the resulting mutation density is typically low compared to a mutagen, which causes predominantly single nucleotide variations. Furthermore, large genomic rearrangements might impede or even prevent the identification of the causal mutation by breaking apart a set of linked genes.

Biological material. Mapping-by-sequencing (MBS) can be based on four different sets of biological material. A classical mapping population scheme was frequently used during the first MBS experiments. This involved outcrossing of mutagenized plants with diverged strains followed by one round of selfing to generate the mapping population [147,148]. Sequencing was performed on two genomic F2 pools of mutant and wildtype plants, respectively. Starting with *A. thaliana*, this method was rapidly applied to other model organisms [149,150]. An isogenic population is generated by crossing homozygous mutants with the non-mutagenized progenitor, resulting in segregation of subtle phenotypic differences in the F2 population [151]. Therefore, the only segregating genetic variation is that induced by mutagens. MBS is performed as described above. Homozygosity mapping uses only the genomes of affected individuals, originally in the context of recessive disease alleles in inbred humans [152]. In order to identify the causal homozygous mutation, the genomes are screened for regions with low heterozygosity. This approach enables MBS for species where a generation of a mapping population is not feasible [152,153] and no prior knowledge about the parental alleles [154] or crossing history is needed [155]. Sequencing of individual mutant genomes [144] is an expensive, but even more powerful approach. Phenotyping errors can contaminate pools in MBS, but this approach allows an *in silico* pooling.

Resolution and accuracy. In general, correct phenotyping of each individual of the mapping population is essential for the accuracy of MBS approaches. Contamination of the mapping population with incorrectly phenotyped individuals results in a larger mapping interval, thus complicating the identification of the causal mutation [156]. Therefore, the resolution of MBS depends on the sampling size of correctly phenotyped and genotyped individuals in the mapping population [144]. However, the resolution is only slightly affected by the number of backcrossed generations [157]. As with conventional methods (e.g. classic genetic markers), re-sequencing data can be used to

fine map the trait(s) of interest in a crossing population [158]. The higher the number of recombinants analyzed, the narrower the final mapping interval. All variants can be considered as markers and thus the variant with the closest link to the trait hints towards the genomic position of the underlying locus. Due to the high marker density derived from natural polymorphisms in the recombinant mapping population, a stringent marker selection decreases the number of false-positive markers. However, at the same time the risk of excluding causal mutations increases, leading to a critical trade-off.

Mapping-by-sequencing applications. SHOREmap demonstrated the applicability of MBS in *A. thaliana* [144,147]. Following projects applied MBS to various crop species including sugar beet [159], rice [151], maize [160], barley [161], and cotton [162]. Liu *et al*. applied a modification of MBS to maize for the identification of a drought tolerance locus: BSR-Seq [160]. BSR-Seq uses RNA-Seq reads for the identification of causal mutations without any prior knowledge about polymorphic markers. As a proof of concept, RNA-Seq was performed for the recessive *glossy3 (gl3)* mutation in a segregating F2 population. The *gl3* gene encodes a putative R2R3 type *myb* transcription factor, which regulates the biosynthesis of very-long-chain fatty acids, which are precursors of epicuticular waxes. Rice seedlings lacking *glossy3* show an extremely thick epicuticular wax on juvenile leaves. By using this alternative MBS approach the *gl3* locus was mapped to an interval of approximately 2 Mb. In summary, mapping-by-sequencing is a powerful technique, which will lead to (crop) plants that are well adapted to biotic and abiotic stresses in the future.

Transcriptomics

RNA-Seq

RNA-Seq, the sequencing of cDNAs, emerged as a valuable method for 1) gene expression analysis, 2) *de novo* transcriptome assembly, and 3) the generation of hints for the gene annotation. The Illumina sequencing workflow of cDNA is very similar to the sequencing of genomic DNA. Besides RNA-Seq, the direct sequencing of RNA became broadly available with ONT sequencing [163]. In addition, PacBio provides Iso-Seq to reveal the sequence of full length transcripts, which can facilitate gene annotation in plants [164].

Gene expression analysis. Short RNA-Seq reads replaced previous methods for systematic gene expression analyses like microarrays almost completely [165–167]. Without any prior knowledge about the sequence, the abundance of transcripts can be quantified [165,168,169], e.g. by generating a *de novo* transcriptome assembly based on the RNA-Seq reads (see below) [170,171]. RNA-Seq even allows to distinguish between different transcript isoforms of the same gene [165,168,169]. Saturation of the signal as observed for microarrays is no longer an issue as the number of reads is proportional to the transcript abundance [165,167]. Low amounts of samples can be analyzed and transcripts with low abundance can be detected, because a single read would be sufficient to reveal the presence of a certain transcript [165,172]. Transcript quantification can be performed based on alignments against a reference sequence, e.g. using STAR [88], or alignment-free, e.g. via Kallisto [173] (Table 6). Information about the transcript abundance can be subjected to downstream analysis like the identification of differentially expressed genes between samples e.g. via DESeq2 [174]. An alternative approach is the identification of co-expressed genes or the construction of co-expression networks as described in [175] and references therein.

Table 6: RNA-Seq gene expression tools. This table is an **i**ncomplete list of tools related to RNA-Seq analyses. Some tools allow the quantification of transcript abundances, while others are involved in the statistical analysis of the resulting abundance values.

Name	Availability	Link	Function	Reference
featureCounts	Binary available	http://bioinf.wehi.edu.au/featureCounts/	Read counting	[176]
HTSeq	Installation required	https://htseq.readthedocs.io/en/release_0.11.1/	Read counting	[177]
Kallisto	Installation required	https://pachterlab.github.io/kallisto/about		[173]
DESeq2	R package	https://www.bioconductor.org/packages//2.12/bioc/html/DESeq2.html	Differential gene expression analysis	[174]
Limma	R package	https://bioconductor.org/packages/release/bioc/html/limma.html	Differential gene expression analysis	[178]
PIANO	R package	https://bioconductor.org/packages/release/bioc/html/piano.html	GO / pathway enrichment analysis	[179]
WEGO	Online	http://wego.genomics.org.cn/	GO enrichment analysis	[180]
gProfiler	Online	https://biit.cs.ut.ee/gprofiler/gost	GO enrichment analysis	[181]
Mercator	Online	https://www.plabipd.de/portal/web/guest/mercator4	Pathway analysis	[182]
MapMan	Online	https://plabipd.de/portal/mapman	Pathway analysis	[182]
BioMart	Online	http://plants.ensembl.org/biomart/martview	Pathway analysis	[183]
Plant Reactome	Online	https://plantreactome.gramene.org	Pathway analysis	[184]

***De novo* transcriptome assembly**. RNA-Seq reads contain comprehensive information about the transcript sequences. Therefore, a *de novo* assembly can be generated to reveal the sequences of transcripts present in the analyzed sample [185]. *De novo* transcriptome assemblies were frequently applied to discover candidate genes which are responsible for a certain trait of interest [170,186,187]. One of the most popular transcriptome assemblers is Trinity [188] which comprises three sequentially applied modules. Trinity performs an *in silico* normalization of the provided reads, i.e. identical reads are filtered out to achieve a similar coverage depth for all transcripts. Supplying stranded RNA-Seq reads, i.e. reads originating from a specified strand, enables to distinguish between reads originating from mRNAs and reads originating from regulatory antisense transcripts. Trinity performed well in benchmarking studies [189,190], but there are more tools that can be evaluated on a given data set (Table 7). Several transcriptome assemblers including Cufflinks [191], Trinity [188], and StringTie [192] allow the integration of a genome sequence for reference-based or genome-guided assembly.

After generation of an initial assembly, very short sequences as well as bacterial and fungal contamination sequences are usually filtered out based on sequence similarity to databases. Since no introns are included in assembled transcript sequences, the identification of protein coding regions can be performed by searching for open reading frames of sufficient length. ORFfinder [193], OrfPredictor [194], and Transdecoder [188] can perform this task. Collapsing very similar sequences is sometimes required and can be performed by CD-HIT [195,196]. Once a final set of sequences is identified, the assignment of a functional annotation is usually the next step. Sequence similarity to functionally annotated databases like swissprot [197,198] can be harnessed to transfer the functional annotation to the newly assembled sequences. InterProScan5 [199] assigns functional annotations including gene ontology (GO) terms and identifies Pfam domains.

Table 7: *De novo* transcriptome assembly tools. This table is an **i**ncomplete list of tools which can be applied to generate plant transcriptome assemblies based on RNA-Seq data.

Name	Availability	Link	Reference
Trinity	Installation required	https://github.com/trinityrnaseq/trinityrnaseq	[188]
rnaSPAdes	Binary available	http://cab.spbu.ru/software/rnaspades	[200]
SPAdes	Binary available	http://cab.spbu.ru/software/spades	[201]
Trans-ABySS	Installation required	https://github.com/bcgsc/transabyss	[202]
Bridger	Installation required	https://github.com/fmaguire/Bridger_Assembler	[203]
SOAPdenovo-Trans	Installation required	https://github.com/aquaskyline/SOAPdenovo-Trans	[204]
Oases	Installation required	https://github.com/dzerbino/oases	[205]
IDBA-Tran	Installation required	https://github.com/loneknightpy/idba	[206]
BinPacker	Installation required	https://github.com/macmanes-lab/BinPacker	[207]
Shannon	Installation required	https://github.com/sreeramkannan/Shannon	[208]

Gene prediction hints. Since RNA-Seq reads reveal transcript sequences, they can be incorporated in the prediction of genes. The alignment of RNA-Seq reads to a genome assembly indicates the positions of introns through gaps in the alignment. In addition, continuously aligned parts of RNA-Seq reads reveal exon positions. STAR [88] and HISAT2 [124] are suitable tools for the mapping of RNA-Seq reads. If reads are already assembled into contigs, exonerate [90] could be utilized to align transcript sequences to an assembly. Dedicated alignment tools also allow the incorporation of peptide sequences as hints by aligning the sequences of well annotated species against the new assembly. Examples for such peptide alignment tools are exonerate [90] and BLAT [87].

Future directions

Recent developments in sequencing technologies enabled the cost-efficient generation of genome and transcriptome sequences for numerous plant species of interest [19,20]. Most of the traditional plant research already benefits from the availability of sequence information for the respective species of interest. This technological progress enables completely new research projects like comparative genomics of large taxonomic groups. Re-sequencing projects, which rely on a reference sequence for comparison, might be replaced by independent *de novo* genome assemblies for all samples of interest [20].

The availability of large sequence data-sets will also lead to more data-based studies which just re-use the existing sequence data-sets. These publicly available data-sets can be harnessed to answer novel questions which could not have been addressed before [85].

Availability of plant genome sequences can foster the research on and usage of orphan crops [209] and help during *de novo* domestication of crops [210]. Intensifying research activity in this field is especially important to cope with global warming and climatic changes.

Acknowledgements. We thank Jens Theine and Andreas Rempel for helpful comments on the manuscript.

References

1. Somssich M. A short history of *Arabidopsis thaliana* (L.) Heynh. Columbia-0 [Internet]. PeerJ Inc.; 2018 Sep. Report No.: e26931v4. doi:10.7287/peerj.preprints.26931v4

2. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11: 31–46. doi:10.1038/nrg2626

3. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet. 2004;5: 335–344. doi:10.1038/nrg1325

4. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26: 1135–1145. doi:10.1038/nbt1486

5. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19: R227–R240. doi:10.1093/hmg/ddq416

6. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11: 759–769. doi:10.1111/j.1755-0998.2011.03024.x

7. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13: 341. doi:10.1186/1471-2164-13-341

8. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351. doi:10.1038/nrg.2016.49

9. Peterson DG, Arick M. Sequencing Plant Genomes. 2018; 1–85. doi:10.1007/124_2018_18

10. Metzker ML. Sequencing in real time. Nat Biotechnol. 2009;27: 150–151. doi:10.1038/nbt0209-150

11. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959

12. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9: 387–402. doi:10.1146/annurev.genom.9.081307.164359

13. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94: 441–448.

14. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.

15. Li C, Lin F, An D, Wang W, Huang R. Genome Sequencing and Assembly by Long Reads in Plants. Genes. 2017;9. doi:10.3390/genes9010006

16. McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D, et al. Illumina TruSeq Synthetic Long-Reads Empower *De Novo* Assembly and Resolve Complex, Highly-Repetitive Transposable Elements. PLOS ONE. 2014;9: e106689. doi:10.1371/journal.pone.0106689

17. Payne A, Holmes N, Rakyan V, Loose M. Whale watching with BulkVis: A graphical viewer for Oxford Nanopore bulk fast5 files. bioRxiv. 2018; 312256. doi:10.1101/312256

18. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36: 338–345. doi:10.1038/nbt.4060

19. Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF. Plant genome sequencing — applications for crop improvement. Curr Opin Biotechnol. 2014;26: 31–37. doi:10.1016/j.copbio.2013.08.019

20. Jiao W-B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017;36: 64–70. doi:10.1016/j.pbi.2017.02.002

21. Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, et al. The Sequenced Angiosperm Genomes and Genome Databases. Front Plant Sci. 2018;9. doi:10.3389/fpls.2018.00418

22. Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, et al. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci Rep. 2018;8: 10931. doi:10.1038/s41598-018-29334-5

23. Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA, et al. Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. GigaScience. 2018;7. doi:10.1093/gigascience/giy033

24. Hu Y, Green GS, Milgate AW, Stone EA, Rathjen JP, Schwessinger B. Pathogen Detection and Microbiome Analysis of Infected Wheat Using a Portable DNA Sequencer. Phytobiomes J. 2019; PBIOMES-01-19-0004-R. doi:10.1094/PBIOMES-01-19-0004-R

25. Stoiber M, Brown J. BasecRAWller: Streaming Nanopore Basecalling Directly from Raw Signal. bioRxiv. 2017; 133058. doi:10.1101/133058

26. Nguyen SH, Duarte TPS, Coin LJM, Cao MD. Real-time demultiplexing Nanopore barcoded sequencing data with npBarcode. Bioinformatics. 2017;33: 3988–3990. doi:10.1093/bioinformatics/btx537

27. Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data [Internet]. 2010 [cited 14 Dec 2017]. Available: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

28. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics. 2016;32: 3047–3048. doi:10.1093/bioinformatics/btw354

29. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170

30. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet TIG. 2008;24: 133–141. doi:10.1016/j.tig.2007.12.007

31. Chaisson MJ, Brinza D, Pevzner PA. *De novo* fragment assembly with short mate-paired reads: Does the read length matter? Genome Res. 2009;19: 336–346. doi:10.1101/gr.079053.108

32. Myers JEW. A history of DNA sequence assembly. It - Inf Technol. 2016;58: 126–132. doi:10.1515/itit-2015-0047

33. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. GigaScience. 2012;1: 18. doi:10.1186/2047-217X-1-18

34. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci. 2011;108: 1513–1518. doi:10.1073/pnas.1017351108

35. Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M, et al. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. Genome Res. 2014;24: 1384–1395. doi:10.1101/gr.170720.113

36. QIAGEN. QIAGEN Bioinformatics - Sample to Insight. In: QIAGEN Bioinformatics [Internet]. 2016 [cited 16 Dec 2018]. Available: https://www.qiagenbioinformatics.com/

37. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. GigaScience. 2013;2: 10. doi:10.1186/2047-217X-2-10

38. Chikhi R, Medvedev P. Informed and automated k-mer size selection for genome assembly. Bioinformatics. 2014;30: 31–37. doi:10.1093/bioinformatics/btt310

39. Shariat B, Movahedi NS, Chitsaz H, Boucher C. HyDA-Vista: towards optimal guided selection of k-mer size for sequence assembly. BMC Genomics. 2014;15: S9. doi:10.1186/1471-2164-15-S10-S9

40. Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S, et al. GAGE: A critical evaluation of genome assemblies and assembly algorithms. Genome Res. 2012;22: 557–567. doi:10.1101/gr.131383.111

41. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics. 2011;27: 578–579. doi:10.1093/bioinformatics/btq683

42. Boetzer M, Pirovano W. Toward almost closed genomes with GapFiller. Genome Biol. 2012;13: R56. doi:10.1186/gb-2012-13-6-r56

43. Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, Birol I. Sealer: a scalable gap-closing application for finishing draft genomes. BMC Bioinformatics. 2015;16. doi:10.1186/s12859-015-0663-4

44. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A *De Novo* Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11: e0164321. doi:10.1371/journal.pone.0164321

45. Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A, et al. TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. bioRxiv. 2019; 631648. doi:10.1101/631648

46. Zerbino DR, Birney E. Velvet: Algorithms for *de novo* short read assembly using de Bruijn graphs. Genome Res. 2008;18: 821–829. doi:10.1101/gr.074492.107

47. Boisvert S, Laviolette F, Corbeil J. Ray: Simultaneous Assembly of Reads from a Mix of High-Throughput Sequencing Technologies. J Comput Biol. 2010;17: 1519–1533. doi:10.1089/cmb.2009.0238

48. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinforma Oxf Engl. 2013;29: 2669–2677. doi:10.1093/bioinformatics/btt476

49. Simpson JT, Durbin R. Efficient *de novo* assembly of large genomes using compressed data structures. Genome Res. 2012;22: 549–556. doi:10.1101/gr.126953.111

50. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat Commun. 2018;9: 541. doi:10.1038/s41467-018-03016-2

51. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, et al. A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set. PLOS ONE. 2019;14: e0216233. doi:10.1371/journal.pone.0216233

52. Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs KL, Eguiarte LE, et al. Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. Proc Natl Acad Sci. 2017;114: 12003–12008. doi:10.1073/pnas.1706367114

53. Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ. Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (Amaranthus hypochondriacus) chromosomes provide insights into genome evolution. BMC Biol. 2017;15: 74. doi:10.1186/s12915-017-0412-4

54. Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods. 2016;13: 1050–1054. doi:10.1038/nmeth.4035

55. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017; gr.215087.116. doi:10.1101/gr.215087.116

56. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37: 540. doi:10.1038/s41587-019-0072-8

57. Li H. Minimap and miniasm: fast mapping and *de novo* assembly for noisy long sequences. Bioinforma Oxf Engl. 2016;32: 2103–2110. doi:10.1093/bioinformatics/btw152

58. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. bioRxiv. 2019; 530972. doi:10.1101/530972

59. Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio RD, et al. The *Cardamine hirsuta* genome offers insight into the evolution of morphological diversity. Nat Plants. 2016;2: 16167. doi:10.1038/nplants.2016.167

60. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome Res. 2017; gr.213652.116. doi:10.1101/gr.213652.116

61. Lin HC, Goldstein S, Mendelowitz L, Zhou S, Wetzel J, Schwartz DC, et al. AGORA: Assembly Guided by Optical Restriction Alignment. BMC Bioinformatics. 2012;13: 189. doi:10.1186/1471-2105-13-189

62. Tang H, Lyons E, Town CD. Optical mapping in plant comparative genomics. GigaScience. 2015;4. doi:10.1186/s13742-015-0044-y

63. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31: 1119–1125. doi:10.1038/nbt.2727

64. Phillippy AM. New advances in sequence assembly. Genome Res. 2017;27: xi–xiii. doi:10.1101/gr.223057.117

65. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinformatics. 2012;13: 238. doi:10.1186/1471-2105-13-238

66. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinforma Oxf Engl. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

67. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. Nat Methods. 2015;12: 733–735. doi:10.1038/nmeth.3444

68. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio. 2013; Available: http://arxiv.org/abs/1303.3997

69. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLOS ONE. 2014;9: e112963. doi:10.1371/journal.pone.0112963

70. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017;27: 737–746. doi:10.1101/gr.214270.116

71. Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD. REAPR: a universal tool for genome assembly evaluation. Genome Biol. 2013;14: R47. doi:10.1186/gb-2013-14-5-r47

72. Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ. KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. Bioinformatics. 2017;33: 574–576. doi:10.1093/bioinformatics/btw663

73. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinforma Oxf Engl. 2015;31: 3210–3212. doi:10.1093/bioinformatics/btv351

74. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, et al. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019;47: D807–D811. doi:10.1093/nar/gky1053

75. Dohmen E, Kremer LPM, Bornberg-Bauer E, Kemena C. DOGMA: domain-based transcriptome and proteome quality assessment. Bioinformatics. 2016;32: 2577–2581. doi:10.1093/bioinformatics/btw231

76. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: *ab initio* prediction of alternative transcripts. Nucleic Acids Res. 2006;34: W435–W439. doi:10.1093/nar/gkl200

77. Hoff KJ, Stanke M. Predicting Genes in Single Genomes with AUGUSTUS. Curr Protoc Bioinforma. 2019;65: e57. doi:10.1002/cpbi.57

78. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33: 6494–6506. doi:10.1093/nar/gki937

79. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Res. 2008;18: 1979–1990. doi:10.1101/gr.081612.108

80. Borodovsky M, Lomsadze A. Eukaryotic Gene Prediction Using GeneMark.hmm-E and GeneMark-ES. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. 2011;CHAPTER: Unit-4.610. doi:10.1002/0471250953.bi0406s35

81. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;42: e119. doi:10.1093/nar/gku557

82. Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. Mol Biol Evol. 2018;35: 543–548. doi:10.1093/molbev/msx319

83. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;44: e89. doi:10.1093/nar/gkw092

84. Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics. 2018;19: 189. doi:10.1186/s12859-018-2203-5

85. Pucker B, Brockington SF. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. BMC Genomics. 2018;19: 980. doi:10.1186/s12864-018-5360-z

86. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. BMC Res Notes. 2017;10. doi:https://doi.org/10.1186/s13104-017-2985-y

87. Kent WJ. BLAT—The BLAST-Like Alignment Tool. Genome Res. 2002;12: 656–664. doi:10.1101/gr.229202

88. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

89. Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinforma. 2015;51: 11.14.1-11.14.19. doi:10.1002/0471250953.bi1114s51

90. Slater GS, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6: 31–31. doi:10.1186/1471-2105-6-31

91. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12: 491. doi:10.1186/1471-2105-12-491

92. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;32: 767–769. doi:10.1093/bioinformatics/btv661

93. Souvorov A, Kapustin Y, Kiryutin B, Chetvernin V, Tatusova T, Lipman D. Gnomon – NCBI eukaryotic gene prediction tool. 2010; Available: http://www.ncbi.nlm.nih.gov/core/assets/genome/files/Gnomon-description.pdf

94. Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, et al. Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol. 2002;3: research0029.1. doi:10.1186/gb-2002-3-6-research0029

95. Fickett JW, Hatzigeorgiou AG. Eukaryotic Promoter Recognition. Genome Res. 1997;7: 861–878. doi:10.1101/gr.7.9.861

96. Salzberg SL. Next-generation genome annotation: we still struggle to get it right. Genome Biol. 2019;20: 92. doi:10.1186/s13059-019-1715-2

97. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29: 2933–2935. doi:10.1093/bioinformatics/btt509

98. Chan PP, Lin BY, Mak AJ, Lowe TM. tRNAscan-SE 2.0: Improved Detection and Functional Classification of Transfer RNA Genes. bioRxiv. 2019; 614032. doi:10.1101/614032

99. Smit A, Hubley R, Green P. RepeatMasker Frequently Open-4.0 [Internet]. 2015. Available: http://www.repeatmasker.org/

100. Bayer PE, Edwards D, Batley J. Bias in resistance gene prediction due to repeat masking. Nat Plants. 2018;4: 762. doi:10.1038/s41477-018-0264-0

101. Michael TP. Plant genome size variation: bloating and purging DNA. Brief Funct Genomics. 2014;13: 308–317. doi:10.1093/bfgp/elu005

102. Vicient CM, Casacuberta JM. Impact of transposable elements on polyploid plant genomes. Ann Bot. 2017;120: 195–207. doi:10.1093/aob/mcx078

103. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element Diversification in *De Novo* Annotation Approaches. PLoS ONE. 2011;6. doi:10.1371/journal.pone.0016526

104. El Baidouri M, Kim KD, Abernathy B, Arikit S, Maumus F, Panaud O, et al. A new approach for annotation of transposable elements using small RNA mapping. Nucleic Acids Res. 2015;43: e84–e84. doi:10.1093/nar/gkv257

105. Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C, et al. A call for benchmarking transposable element annotation methods. Mob DNA. 2015;6. doi:10.1186/s13100-015-0044-6

106. Price AL, Jones NC, Pevzner PA. *De novo* identification of repeat families in large genomes. Bioinforma Oxf Engl. 2005;21 Suppl 1: i351-358. doi:10.1093/bioinformatics/bti1018

107. Estill JC, Bennetzen JL. The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. Plant Methods. 2009;5: 8. doi:10.1186/1746-4811-5-8

108. Saha S, Bridges S, Magbanua ZV, Peterson DG. Computational Approaches and Tools Used in Identification of Dispersed Repetitive DNA Sequences. Trop Plant Biol. 2008;1: 85–96. doi:10.1007/s12042-007-9007-5

109. Bergman CM, Quesneville H. Discovering and detecting transposable elements in genome sequences. Brief Bioinform. 2007;8: 382–392. doi:10.1093/bib/bbm048

110. Lerat E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity. 2010;104: 520–533. doi:10.1038/hdy.2009.165

111. Kemena C, Dohmen E, Bornberg-Bauer E. DOGMA: a web server for proteome and transcriptome quality assessment. Nucleic Acids Res. doi:10.1093/nar/gkz366

112. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5: 59. doi:10.1186/1471-2105-5-59

113. Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM, et al. 1,135 Genomes Reveal the Global Pattern of Polymorphism in *Arabidopsis thaliana*. Cell. 2016;166: 481–491. doi:10.1016/j.cell.2016.05.063

114. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9: 357–359. doi:10.1038/nmeth.1923

115. Broad Institute. Picard toolkit [Internet]. Broad Institute; 2019. Available: https://github.com/broadinstitute/picard

116. Marco-Sola S, Sammeth M, Guigó R, Ribeca P. The GEM mapper: fast, accurate and versatile alignment by filtration. Nat Methods. 2012;9: 1185–1188. doi:10.1038/nmeth.2221

117. Bushnell B. BBMap - Browse Files at SourceForge.net [Internet]. [cited 28 May 2019]. Available: https://sourceforge.net/projects/bbmap/files/

118. NovoCraft. NovoAlign [Internet]. 2010 [cited 27 May 2019]. Available: http://www.novocraft.com/products/novoalign/

119. Sedlazeck FJ, Rescheneder P, von Haeseler A. NextGenMap: fast and accurate read mapping in highly polymorphic genomes. Bioinformatics. 2013;29: 2790–2791. doi:10.1093/bioinformatics/btt468

120. Li H. MAQ [Internet]. [cited 28 May 2019]. Available: http://maq.sourceforge.net/maq-manpage.shtml#12

121. Smith AD, Chung W-Y, Hodges E, Kendall J, Hannon G, Hicks J, et al. Updates to the RMAP short-read mapping software. Bioinformatics. 2009;25: 2841–2842. doi:10.1093/bioinformatics/btp533

122. Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT. MOSAIK: A Hash-Based Algorithm for Accurate Next-Generation Sequencing Short-Read Mapping. PLOS ONE. 2014;9: e90581. doi:10.1371/journal.pone.0090581

123. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, et al. Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. PLOS Comput Biol. 2009;5: e1000502. doi:10.1371/journal.pcbi.1000502

124. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12: 357–360. doi:10.1038/nmeth.3317

125. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20: 1297–1303. doi:10.1101/gr.107524.110

126. Auwera GAV der, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A, et al. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. Curr Protoc Bioinforma. 2013;43: 11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

127. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

128. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44: e108. doi:10.1093/nar/gkw227

129. Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N. Fast and sensitive mapping of nanopore sequencing reads with GraphMap. Nat Commun. 2016;7: 11307. doi:10.1038/ncomms11307

130. Jain M, Fiddes I, Miga KH, Olsen HE, Paten B, Akeson M. Improved data analysis for the MinION nanopore sequencer. Nat Methods. 2015;12: 351–356. doi:10.1038/nmeth.3290

131. Szalay T, Golovchenko JA. *De novo* sequencing and variant calling with nanopores using PoreSeq. Nat Biotechnol. 2015;33: 1087–1091. doi:10.1038/nbt.3360

132. Heller D, Vingron M. SVIM: Structural Variant Identification using Mapped Long Reads. bioRxiv. 2018; 494096. doi:10.1101/494096

133. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36: 983–987. doi:10.1038/nbt.4235

134. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 2011;39: e132–e132. doi:10.1093/nar/gkr599

135. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22: 568–576. doi:10.1101/gr.129684.111

136. Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH, et al. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. Nucleic Acids Res. 2012;40: 11189–11201. doi:10.1093/nar/gks918

137. Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wgs500 Consortium, et al. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. Nat Genet. 2014;46: 912–918. doi:10.1038/ng.3036

138. Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, et al. SNP detection for massively parallel whole-genome resequencing. Genome Res. 2009;19: 1124–1132. doi:10.1101/gr.088013.108

139. Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA, et al. A SNP discovery method to assess variant allele probability from next-generation resequencing data. Genome Res. 2010;20: 273–280. doi:10.1101/gr.096388.109

140. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. ArXiv12073907 Q-Bio. 2012; Available: http://arxiv.org/abs/1207.3907

141. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin). 2012;6: 80–92. doi:10.4161/fly.19695

142. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38: e164. doi:10.1093/nar/gkq603

143. Baasner J-S, Howard D, Pucker B. Influence of neighboring small sequence variants on functional impact prediction. bioRxiv. 2019; 596718. doi:10.1101/596718

144. Schneeberger K. Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet. 2014;15: 662–676. doi:10.1038/nrg3745

145. Garcia V, Bres C, Just D, Fernandez L, Tai FWJ, Mauxion J-P, et al. Rapid identification of causal mutations in tomato EMS populations via mapping-by-sequencing. Nat Protoc. 2016;11: 2401–2418. doi:10.1038/nprot.2016.143

146. Sikora P, Chawade A, Larsson M, Olsson J, Olsson O. Mutagenesis as a Tool in Plant Genetics, Functional Genomics, and Breeding. Int J Plant Genomics. 2011;2011. doi:10.1155/2011/314829

147. Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, et al. SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods. 2009;6: 550–551. doi:10.1038/nmeth0809-550

148. Cuperus JT, Montgomery TA, Fahlgren N, Burke RT, Townsend T, Sullivan CM, et al. Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing. Proc Natl Acad Sci U S A. 2010;107: 466–471. doi:10.1073/pnas.0913203107

149. Wenger JW, Schwartz K, Sherlock G. Bulk Segregant Analysis by High-Throughput Sequencing Reveals a Novel Xylose Utilization Gene from Saccharomyces cerevisiae. PLoS Genet. 2010;6. doi:10.1371/journal.pgen.1000942

150. Leshchiner I, Alexa K, Kelsey P, Adzhubei I, Austin-Tse CA, Cooney JD, et al. Mutation mapping and identification by whole-genome sequencing. Genome Res. 2012;22: 1541–1548. doi:10.1101/gr.135541.111

151. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, et al. Genome sequencing reveals agronomically important loci in rice using MutMap. Nat Biotechnol. 2012;30: 174–178. doi:10.1038/nbt.2095

152. Lander ES, Botstein D. Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science. 1987;236: 1567–1570.

153. Singh R, Leslie Low E-T, Ooi LC-L, Ong-Abdullah M, Chin TN, Nagappan J, et al. The oil palm Shell gene controls oil yield and encodes a homologue of SEEDSTICK. Nature. 2013;500: 340–344. doi:10.1038/nature12356

154. Voz ML, Coppieters W, Manfroid I, Baudhuin A, Von Berg V, Charlier C, et al. Fast Homozygosity Mapping and Identification of a Zebrafish ENU-Induced Mutation by Whole-Genome Sequencing. PLoS ONE. 2012;7. doi:10.1371/journal.pone.0034671

155. Bowen ME, Henke K, Siegfried KR, Warman ML, Harris MP. Efficient Mapping and Cloning of Mutations in Zebrafish by Low-Coverage Whole-Genome Sequencing. Genetics. 2012;190: 1017–1024. doi:10.1534/genetics.111.136069

156. Greenberg MV, Ausin I, Chan SW, Cokus SJ, Cuperus JT, Feng S, et al. Identification of genes required for de novo DNA methylation in Arabidopsis. Epigenetics. 2011;6: 344–354. doi:10.4161/epi.6.3.14242

157. James GV, Patel V, Nordström KJ, Klasen JR, Salomé PA, Weigel D, et al. User guide for mapping-by-sequencing in Arabidopsis. Genome Biol. 2013;14: R61. doi:10.1186/gb-2013-14-6-r61

158. Schneeberger K, Weigel D. Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci. 2011;16: 282–288. doi:10.1016/j.tplants.2011.02.006

159. Ries D, Holtgräwe D, Viehöver P, Weisshaar B. Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. BMC Genomics. 2016;17. doi:10.1186/s12864-016-2566-9

160. Liu S, Yeh C-T, Tang HM, Nettleton D, Schnable PS. Gene Mapping via Bulked Segregant RNA-Seq (BSR-Seq). PLoS ONE. 2012;7. doi:10.1371/journal.pone.0036406

161. Mascher M, Jost M, Kuon J-E, Himmelbach A, Aßfalg A, Beier S, et al. Mapping-by-sequencing accelerates forward genetics in barley. Genome Biol. 2014;15: R78. doi:10.1186/gb-2014-15-6-r78

162. Chen W, Yao J, Chu L, Yuan Z, Li Y, Zhang Y. Genetic mapping of the nulliplex-branch gene (gb_nb1) in cotton using next-generation sequencing. Theor Appl Genet. 2015;128: 539–547. doi:10.1007/s00122-014-2452-2

163. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15: 201–206. doi:10.1038/nmeth.4577

164. Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M, et al. Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. Genome Biol. 2015;16: 184. doi:10.1186/s13059-015-0729-7

165. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63. doi:10.1038/nrg2484

166. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320: 1344–1349. doi:10.1126/science.1158441

167. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5: 621–628. doi:10.1038/nmeth.1226

168. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 2017;89: 789–804. doi:10.1111/tpj.13415

169. Cheng B, Furtado A, Henry RJ. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. GigaScience. 2017;6. doi:10.1093/gigascience/gix086

170. Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J, et al. High Quality *de Novo* Transcriptome Assembly of *Croton tiglium*. Front Mol Biosci. 2018;5. doi:https://doi.org/10.3389/fmolb.2018.00062

171. Müller M, Seifert S, Lübbe T, Leuschner C, Finkeldey R. *De novo* transcriptome assembly and analysis of differential gene expression in response to drought in European beech. PloS One. 2017;12: e0184167. doi:10.1371/journal.pone.0184167

172. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018;9: 619. doi:10.1038/s41467-018-02866-0

173. Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol. 2016;34: 525–527. doi:10.1038/nbt.3519

174. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15. doi:https://doi.org/10.1186/s13059-014-0550-8

175. van Dam S, Võsa U, van der Graaf A, Franke L, de Magalhães JP. Gene co-expression analysis for functional classification and gene–disease predictions. Brief Bioinform. 2017;19: 575–592. doi:10.1093/bib/bbw139

176. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinforma Oxf Engl. 2014;30: 923–930. doi:10.1093/bioinformatics/btt656

177. Anders S, Pyl PT, Huber W. HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;31: 166–169. doi:10.1093/bioinformatics/btu638

178. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43: e47. doi:10.1093/nar/gkv007

179. Väremo L, Nielsen J, Nookaew I. Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. Nucleic Acids Res. 2013;41: 4378–4391. doi:10.1093/nar/gkt111

180. Ye J, Zhang Y, Cui H, Liu J, Wu Y, Cheng Y, et al. WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. Nucleic Acids Res. 2018;46: W71–W75. doi:10.1093/nar/gky400

181. Reimand J, Kull M, Peterson H, Hansen J, Vilo J. g:Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res. 2007;35: W193–W200. doi:10.1093/nar/gkm226

182. Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, Hallab A, et al. MapMan4: A Refined Protein Classification and Annotation Framework Applicable to Multi-Omics Data Analysis. Mol Plant. 2019; doi:10.1016/j.molp.2019.01.003

183. Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J, et al. The BioMart community portal: an innovative alternative to large, centralized data repositories. Nucleic Acids Res. 2015;43: W589–W598. doi:10.1093/nar/gkv350

184. Naithani S, Preece J, D'Eustachio P, Gupta P, Amarasinghe V, Dharmawardhana PD, et al. Plant Reactome: a resource for plant pathways and comparative analysis. Nucleic Acids Res. 2017;45: D1029–D1039. doi:10.1093/nar/gkw932

185. Schliesky S, Gowik U, Weber APM, Bräutigam A. RNA-Seq Assembly - Are We There Yet? Front Plant Sci. 2012;3: 220. doi:10.3389/fpls.2012.00220

186. Wu S, Lei J, Chen G, Chen H, Cao B, Chen C. *De novo* Transcriptome Assembly of Chinese Kale and Global Expression Analysis of Genes Involved in Glucosinolate Metabolism in Multiple Tissues. Front Plant Sci. 2017;8. doi:10.3389/fpls.2017.00092

187. Han Y, Wan H, Cheng T, Wang J, Yang W, Pan H, et al. Comparative RNA-seq analysis of transcriptome dynamics during petal development in *Rosa chinensis*. Sci Rep. 2017;7: 43382. doi:10.1038/srep43382

188. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8: 1494–1512. doi:10.1038/nprot.2013.084

189. Hölzer M, Marz M. De novo transcriptome assembly: A comprehensive cross-species comparison of short-read RNA-Seq assemblers. GigaScience. 2019;8. doi:10.1093/gigascience/giz039

190. Behera S, Voshall A, Deogun JS, Moriyama EN. Performance comparison and an ensemble approach of transcriptome assembly. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). 2017. pp. 2226–2228. doi:10.1109/BIBM.2017.8218005

191. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van MB, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28: 511–515. doi:10.1038/nbt.1621

192. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33: 290–295. doi:10.1038/nbt.3122

193. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, et al. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 2003;31: 28–33.

194. Min XJ, Butler G, Storms R, Tsang A. OrfPredictor: predicting protein-coding regions in EST-derived sequences. Nucleic Acids Res. 2005;33: W677-680. doi:10.1093/nar/gki394

195. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22: 1658–1659. doi:10.1093/bioinformatics/btl158

196. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28: 3150–3152. doi:10.1093/bioinformatics/bts565

197. Bairoch A, Apweiler R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. 2000;28: 45–48.

198. The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Res. 2017;45: D158–D169. doi:10.1093/nar/gkw1099

199. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, et al. InterPro in 2017—beyond protein family and domain annotations. Nucleic Acids Res. 2017;45: D190–D199. doi:10.1093/nar/gkw1107

200. Bushmanova E, Antipov D, Lapidus A, Przhibelskiy AD. rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. bioRxiv. 2018; 420208. doi:10.1101/420208

201. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol. 2012;19: 455–477. doi:10.1089/cmb.2012.0021

202. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. *De novo* assembly and analysis of RNA-seq data. Nat Methods. 2010;7: 909–912. doi:10.1038/nmeth.1517

203. Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D, et al. Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. Genome Biol. 2015;16. doi:10.1186/s13059-015-0596-2

204. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics. 2014;30: 1660–1666. doi:10.1093/bioinformatics/btu077

205. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics. 2012;28: 1086–1092. doi:10.1093/bioinformatics/bts094

206. Peng Y, Leung HCM, Yiu S-M, Lv M-J, Zhu X-G, Chin FYL. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics. 2013;29: i326–i334. doi:10.1093/bioinformatics/btt219

207. Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R, et al. BinPacker: Packing-Based De Novo Transcriptome Assembly from RNA-seq Data. PLoS Comput Biol. 2016;12. doi:10.1371/journal.pcbi.1004772

208. Kannan S, Hui J, Mazooji K, Pachter L, Tse D. Shannon: An Information-Optimal de Novo RNA-Seq Assembler. bioRxiv. 2016; 039230. doi:10.1101/039230

209. Chang Y, Liu H, Liu M, Liao X, Sahu SK, Fu Y, et al. The draft genomes of five agriculturally important African orphan crops. GigaScience. 2018; doi:10.1093/gigascience/giy152

210. Fernie AR, Yan J. De Novo Domestication: An Alternative Route toward New Crops for the Future. Mol Plant. 2019;12: 615–631. doi:10.1016/j.molp.2019.03.016