

# Chapter 19

## Genomics and Transcriptomics Advance in Plant Sciences



**Boas Pucker and Hanna Marie Schilbert**

**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.

**Keywords** 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

---

B. Pucker (✉) · H. M. Schilbert  
Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University, Bielefeld, Germany  
e-mail: [bpucker@cebitec.uni-bielefeld.de](mailto:bpucker@cebitec.uni-bielefeld.de)

© Springer Nature Singapore Pte Ltd. 2019  
S. P. Singh et al. (eds.), *Molecular Approaches in Plant Biology and Environmental Challenges*, Energy, Environment, and Sustainability,  
[https://doi.org/10.1007/978-981-15-0690-1\\_19](https://doi.org/10.1007/978-981-15-0690-1_19)

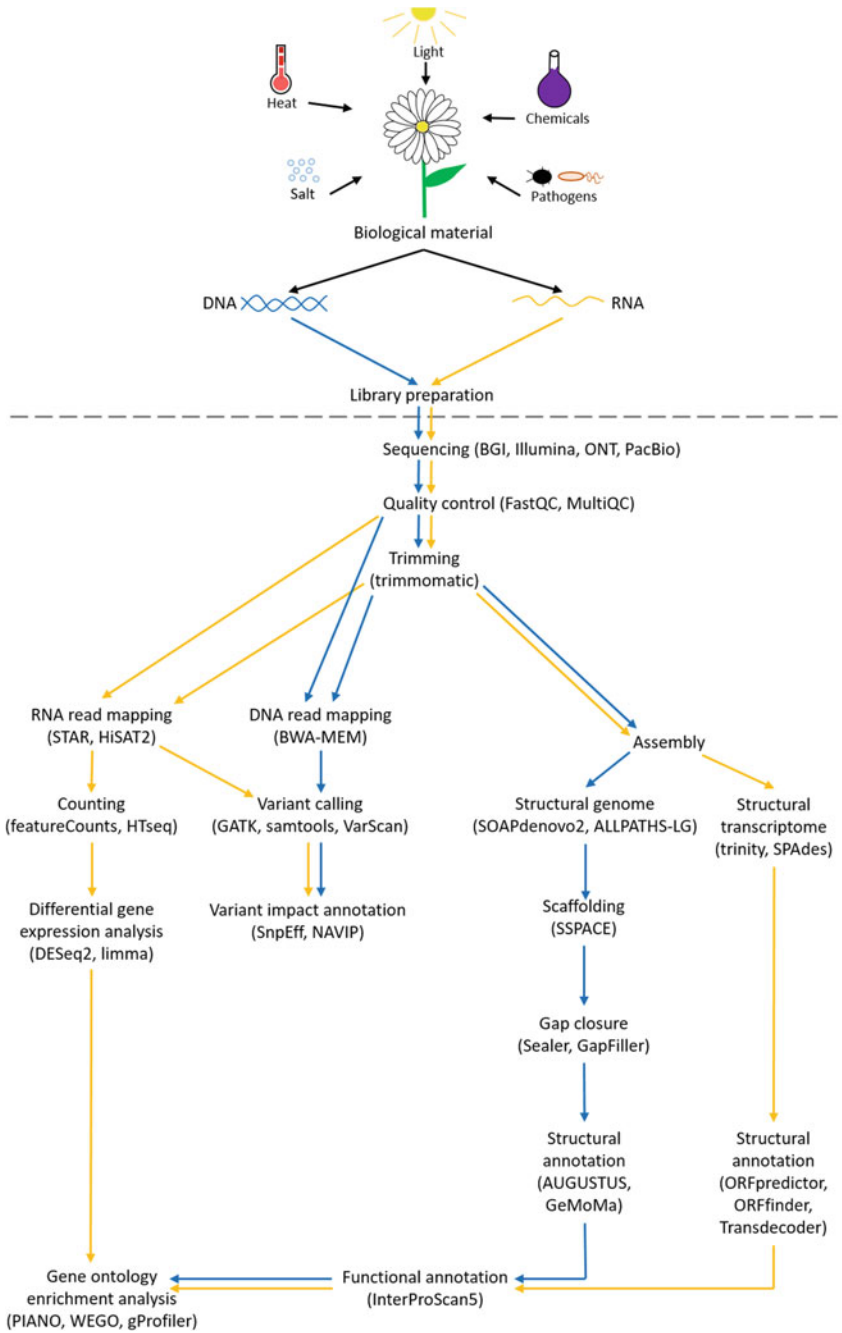
## 19.1 Introduction

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 (Somssich 2018). 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. 19.1).

## 19.2 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 (Metzker 2010; Shendure et al. 2004; Shendure and Ji 2008; Schadt et al. 2010; Glenn 2011; Quail et al. 2012; Goodwin et al. 2016; Peterson and Arick 2018). 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 Biosciences, PacBio) and nanopore sequencing (Oxford Nanopore Technologies, ONT) as third generation sequencing technologies. Technical details of these sequencing technologies were reviewed elsewhere (Metzker 2009, 2010; Shendure and Ji 2008; Quail et al. 2012; Goodwin et al. 2016; Margulies et al. 2005; Mardis 2008a).

Since the invention of chain termination sequencing (Sanger and Coulson 1975; Sanger et al. 1977), substantial technological advances paved the way for cost reductions. Therefore, broad application of high throughput sequencing (Metzker 2010) and more recently long read sequencing technologies (Li et al. 2017) 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



**Fig. 19.1 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

IonTorrent is comparable to Sanger sequencing, but have reduced accuracy. Nevertheless, Illumina has been dominating the market for high throughput sequencing with substantially shorter 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.  $2 \times 300$  nt, can be merged, thus leading to a total length of up to  $<600$  nt. Sophisticated approaches like TrueSeq synthetic long reads (McCoy et al. 2014) 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 as valuable linkage information during genome assembly processes.

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 (Payne et al. 2018) and the longest single read is close to 1 Mbp (Jain et al. 2018). Dropping sequencing costs and the rise of long read technologies enabled sequencing projects for numerous plant species (Bolger et al. 2014a; Jiao and Schneeberger 2017; Chen et al. 2018). 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 (Tyler et al. 2018; Pomerantz et al. 2018). Sequencing in the field opens up opportunities, to monitor pathogens in the field accurately (Hu et al. 2019) and to assess the biodiversity (Pomerantz et al. 2018). Real time base calling and the start of downstream analysis before completion of a sequencing run are beneficial when decisions are time critical (Stoiber and Brown 2017). 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 (Nguyen et al. 2017).

## 19.3 Genomics

### 19.3.1 Genome Assembly

**Quality control and preprocessing** Quality checks via FastQC (Andrews 2010) or MultiQC (Ewels et al. 2016) 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 (Bolger et al. 2014b). 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 (Mardis 2008b; Chaisson et al. 2009; Myers 2016). 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 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 (Luo et al. 2012), ALLPATHS-LG (Gnerre et al. 2011), Platanus (Kajitani et al. 2014), and the proprietary CLC assembler (QIAGEN 2016) are examples for tools which were successfully deployed for the assembly of plant genomes, but there are also many alternatives (Table 19.1). Modification of parameters, especially  $k$ -mer sizes, should be optimized empirically (Bradnam et al. 2013; Chikhi and Medvedev 2014; Shariat et al. 2014; Salzberg et al. 2012). 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 (Boetzer et al. 2011) are available. Next, gaps between contigs within a scaffold can be partially closed, e.g. via Gap-Filler (Boetzer and Pirovano 2012) or Sealer (Paulino et al. 2015). The reduced sequencing costs allowed the assembly of plant genome sequences by single groups (Pucker et al. 2016), but most genome sequences were highly fragmented. More recently, the proprietary NRGene assembler (DeNovoMAGIC™) and the competing open source alternative TRITEX (Monat et al. 2019) are promising substantially improved assemblies.

**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 (Michael et al. 2018; Pucker et al. 2019; Copetti et al. 2017; Lightfoot et al. 2017). 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 19.2).

**Table 19.1** Assembler for second generation sequencing data

Name	Availability	Link	References
CLC	Licence required	<a href="https://www.qiagenbioinformatics.com/products/clc-main-workbench">https://www.qiagenbioinformatics.com/products/clc-main-workbench</a>	QIAGEN (2016)
SOAPdenovo2	Binary available	<a href="https://github.com/aquaskyline/SOAPdenovo2">https://github.com/aquaskyline/SOAPdenovo2</a>	Luo et al. (2012)
Velvet	Installation required	<a href="https://github.com/dzerbino/velvet">https://github.com/dzerbino/velvet</a>	Zerbino and Birney (2008)
ALLPATHS-LG	Installation required	<a href="http://software.broadinstitute.org/allpaths-lg/blog/?page_id=12">http://software.broadinstitute.org/allpaths-lg/blog/?page_id=12</a>	Gnerre et al. (2011)
Ray	Installation required	<a href="http://denovoassembler.sourceforge.net">http://denovoassembler.sourceforge.net</a>	Boisvert et al. (2010)
Newbler	Installation required	<a href="http://sequencing.roche.com">http://sequencing.roche.com</a>	Margulies et al. (2005)
MaSuRCA	Installation required	<a href="https://github.com/alekseyzimin/masurca">https://github.com/alekseyzimin/masurca</a>	Zimin et al. (2013)
SGA	Installation required	<a href="https://github.com/jts/sga">https://github.com/jts/sga</a>	Simpson and Durbin (2012)
Platanus	Installation required	<a href="http://platanus.bio.titech.ac.jp">http://platanus.bio.titech.ac.jp</a>	Kajitani et al. (2014)

This table is an incomplete list of tools that can be applied for the de novo plant genome assembly based on second generation sequencing data

**Table 19.2** Third generation assembler

Name	Availability	Link	References
FALCON	SMRT Link	<a href="https://https://pacb.www.pacb.com/training/smr-link-overview">https://https://pacb.www.pacb.com/training/smr-link-overview</a>	Chin et al. (2016)
Canu	Installation required	<a href="https://https://github.com/marbl/canu">https://https://github.com/marbl/canu</a>	Koren et al. (2017)
Flye	Installation required	<a href="https://github.com/fenderglass/Flye">https://github.com/fenderglass/Flye</a>	Kolmogorov et al. (2019)
Miniasm	Installation required	<a href="https://github.com/lh3/miniasm">https://github.com/lh3/miniasm</a>	Li (2016)
wtdbg2	Installation required	<a href="https://github.com/ruanjue/wtdbg2">https://github.com/ruanjue/wtdbg2</a>	Ruan and Li (2019)

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

FALCON (Chin et al. 2016), Canu (Koren et al. 2017), Flye (Kolmogorov et al. 2019), Miniasm (Li 2016), and wtdbg2 (Ruan and Li 2019) 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 (Pucker et al. 2019; Gan et al. 2016), optical mapping information, e.g. from Bionano Genomics and OptGen (Jiao et al. 2017; Lin et al. 2012; Tang et al. 2015), and Hi-C (Jiao et al. 2017; Burton et al. 2013; Phillippy 2017). Genetic linkage can rely on molecular markers measured in the lab (Pucker et al. 2019) or on sequencing of multiple individual plants of a segregating population by a high throughput method (Gan et al. 2016). 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 (Chaisson and Tesler 2012) and minimap2 (Li 2018). Arrow (Chin et al. 2016) can be applied to polish assemblies based on PacBio reads, while nanopolish (Loman et al. 2015) 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 (Li 2013) is a suitable read mapping tool and Pilon (Walker et al. 2014) 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 (Koren et al. 2017; Vaser et al. 2017). 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.

**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 (Li 2013), and analyzing the distances of paired reads in this mapping, e.g. via REAPR (Hunt et al. 2013). Alternative approaches like implemented in KAT (Mapleson et al. 2017) 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) (Simão

et al. 2015) 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 (Kriventseva et al. 2019). 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 (Dohmen et al. 2016) 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>).

### 19.3.2 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 19.3). AUGUSTUS (Stanke et al. 2006; Hoff and Stanke 2019) and GeneMark derivatives (Lomsadze et al. 2005, 2014; Ter-Hovhannisyan et al. 2008; Borodovsky and Lomsadze 2011) 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 (Waterhouse et al. 2018). In contrast to these *ab initio* approaches, GeMoMa (Keilwagen et al. 2016, 2018) 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 (Pucker and Brockington 2018). 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 (Pucker and Brockington 2018; Pucker et al. 2017). 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,



**Table 19.3** Plant gene prediction tools

Name	Availability	Link	References
AUGUSTUS	Installation required	<a href="https://github.com/Gaius-Augustus/Augustus">https://github.com/Gaius-Augustus/Augustus</a>	Stanke et al. (2006)
BRAKER1	Installation required	<a href="https://github.com/Gaius-Augustus/BRAKER">https://github.com/Gaius-Augustus/BRAKER</a>	Hoff et al. (2016)
GeneMark	Installation required	<a href="http://exon.gatech.edu/GeneMark/license_download.cgi">http://exon.gatech.edu/GeneMark/license_download.cgi</a>	Ter-Hovhannisyan et al. (2008), Borodovsky and Lomsadze (2011), Lomsadze et al. (2014)
GeMoMa	Jar file	<a href="http://www.jstacs.de/index.php/GeMoMa">http://www.jstacs.de/index.php/GeMoMa</a>	Keilwagen et al. (2016, 2018)
Gnomon	Installation required	<a href="ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/+/CURRENT">ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/+/CURRENT</a>	Souvorov et al. (2010)
MAKER2	Registration required	<a href="https://www.yandell-lab.org/software/maker.html">https://www.yandell-lab.org/software/maker.html</a>	Holt and Yandell (2011)
SNAP	Installation required	<a href="https://github.com/KorfLab/SNAP">https://github.com/KorfLab/SNAP</a>	Korf (2004)

This table is an incomplete list of tools that can be applied for gene prediction on plant genome assemblies

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 (Kent 2002) can align long sequences, STAR (Dobin et al. 2013; Dobin and Gingeras 2015) is well suited for the split alignment of RNA-Seq reads. Dedicated tools like exonerate (Slater and Birney 2005) allow the alignment of previously annotated peptide sequences from other species. Resulting alignments can be converted into gene prediction hints. Annotation pipelines like MAKER2 (Holt and Yandell 2011), BRAKER1 (Hoff et al. 2016), and Gnomon (Souvorov et al. 2010) 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 (Pucker et al. 2017; Haas et al. 2002; Fickett and Hatzigeorgiou 1997). 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 (Salzberg 2019).

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 (Nawrocki and Eddy 2013) and tRNAscan-SE2 (Chan et al. 2019) are tools for the prediction of pure RNA genes.

Masking of repeats, e.g. via RepeatMasker (Smit et al. 2015), 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 (Bayer et al. 2018). Although transposable elements and other repeats account

for the major proportion of many plant genomes (Michael 2014; Vicient and Casacuberta 2017), the annotation of repeats is often performed poorly or omitted completely (Flutre et al. 2011; El Baidouri et al. 2015; Hoen et al. 2015). There is a plethora of annotation tools like RepeatScout (Price et al. 2005) and RepeatMasker (Smit et al. 2015). Bioinformatic pipelines were developed to account for weaknesses of single tools and to combine the strengths of many individual tools (Estill and Bennetzen 2009; Saha et al. 2008; Bergman and Quesneville 2007). 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 (Hoen et al. 2015; Lerat 2010). While the annotation of protein encoding genes can be checked for completeness based on BUSCO (Simão et al. 2015) and DOGMA (Dohmen et al. 2016; Kemena et al. 2019), there is no such benchmarking data-set available for TEs.

**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 (Kemena et al. 2019).

### 19.3.3 *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 1135 *A. thaliana* accessions revealed insights into the genomic diversity of this species (Alonso-Blanco et al. 2016). 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 (Li 2013) and bowtie2 (Langmead and Salzberg 2012) are frequently applied tools for the mapping of reads to a reference sequence (Table 19.4). 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 (Broad Institute 2019) 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

**Table 19.4** Read mapping tools

Name	Availability	Link	DNA/RNA	References
BWA-MEM	Installation required	<a href="https://github.com/lh3/bwa">https://github.com/lh3/bwa</a>	DNA	Li (2013)
Bowtie 2	Installation required	<a href="https://github.com/BenLangmead/bowtie2">https://github.com/BenLangmead/bowtie2</a>	DNA	Langmead and Salzberg (2012)
GEM 3	Installation required	<a href="https://github.com/smarco/gem3-mapper">https://github.com/smarco/gem3-mapper</a>	DNA	Marco-Sola et al. (2012)
bbmap	Jar file available	<a href="https://sourceforge.net/projects/bbmap">https://sourceforge.net/projects/bbmap</a>	DNA	Bushnell (2019)
Novoalign	Trial available	<a href="http://www.novocraft.com/products/novoalign">http://www.novocraft.com/products/novoalign</a>	DNA	NovoCraft (2010)
NextGenMap	Installation required	<a href="https://github.com/Cibiv/NextGenMap">https://github.com/Cibiv/NextGenMap</a>	DNA	Sedlazeck et al. (2013)
MAQ	Installation required	<a href="http://maq.sourceforge.net/maq-man.shtml">http://maq.sourceforge.net/maq-man.shtml</a>	DNA	Li (2019)
RMAP	Installation required	<a href="https://github.com/smithlabcode/rmap">https://github.com/smithlabcode/rmap</a>	DNA	Smith et al. (2009)
MOSAİK	Installation required	<a href="https://github.com/waninglee/MOSAİK">https://github.com/waninglee/MOSAİK</a>	DNA	Lee et al. (2014)
segemehl	Installation required	<a href="https://www.bioinf.uni-leipzig.de/Software/segemehl">https://www.bioinf.uni-leipzig.de/Software/segemehl</a>	RNA	Hoffmann et al. (2009)
STAR	Installation required	<a href="https://github.com/alexdobin/STAR">https://github.com/alexdobin/STAR</a>	RNA	Dobin et al. (2013)
HISAT2	Binary available	<a href="https://ccb.jhu.edu/software/hisat2/manual.shtml">https://ccb.jhu.edu/software/hisat2/manual.shtml</a>	RNA	Kim et al. (2015)

This table is an incomplete 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

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.

There are numerous tools for the detection of genomic differences based on a short read mapping (Table 19.5). Genome Analysis Tool Kit (GATK) (McKenna et al. 2010; der Auwera et al. 2013), samtools/bcftools (Li et al. 2009a), and VarDict

**Table 19.5** Variant callers

Name	Availability	Link	Variants	References
DeepVariant	Installation required	<a href="https://github.com/google/deepvariant">https://github.com/google/deepvariant</a>	Small	Poplin et al. (2018)
GATK	Jar file	<a href="https://software.broadinstitute.org/gatk/download">https://software.broadinstitute.org/gatk/download</a>	Small	McKenna et al. (2010), der Auwera et al. (2013)
SNVer	Installation required	<a href="http://snver.sourceforge.net">http://snver.sourceforge.net</a>	Small	Wei et al. (2011)
SAMtools	Jar file	<a href="http://samtools.sourceforge.net">http://samtools.sourceforge.net</a>	Small	Li et al. (2009a)
VarDict	Installation required	<a href="https://github.com/AstraZeneca-NGS/VarDict">https://github.com/AstraZeneca-NGS/VarDict</a>	Small	Lai et al. (2016)
VarScan 2	Jar file	<a href="http://varscan.sourceforge.net">http://varscan.sourceforge.net</a>	Small	Koboldt et al. (2012)
LoFreq	Binary available	<a href="https://csb5.github.io/lofreq/installation">https://csb5.github.io/lofreq/installation</a>	Small	Wilm et al. (2012)
Platypus	Installation required	<a href="https://github.com/andyrimmer/Platypus">https://github.com/andyrimmer/Platypus</a>	Small	Rimmer et al. (2014)
SOAPSnp	Installation required	<a href="https://sourceforge.net/projects/soapsnp">https://sourceforge.net/projects/soapsnp</a>	Small	Li et al. (2009b)
Atlas-SNP2	Installation required	<a href="https://sourceforge.net/projects/atlas2">https://sourceforge.net/projects/atlas2</a>	Small	Shen et al. (2010)
FreeBayes	Installation required	<a href="https://github.com/ekg/freebayes">https://github.com/ekg/freebayes</a>	Small	Garrison and Marth (2012)
SVIM	Installation required	<a href="https://github.com/eldariont/svim">https://github.com/eldariont/svim</a>	Large	Heller and Vingron (2018)
marginAlign	Installation required	<a href="https://github.com/benedictpaten/marginAlign">https://github.com/benedictpaten/marginAlign</a>	Large	Jain et al. (2015)
GraphMap	Installation required	<a href="https://github.com/isovic/graphmap">https://github.com/isovic/graphmap</a>	Large	Sović et al. (2016)
PoreSeq	Installation required	<a href="https://github.com/tszalay/poreseq">https://github.com/tszalay/poreseq</a>	Large	Szalay and Golovchenko (2015)

This table is an incomplete 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

(Lai et al. 2016) 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 (Sović et al. 2016), marginAlign (Jain et al. 2015), and PoreSeq (Szalay and Golovchenko 2015) can align long reads to a reference sequence to call variants. Other tools like SVIM (Heller and Vingron 2018) rely on alignments generated by dedicated long read aligners like minimap2 (Li 2018) or BLASR (Chaisson and Tesler 2012). Identified variants can be subjected to downstream filtering; e.g. based on the number of supporting and contradicting reads.

Once the variants are identified, it is possible to assign functional annotations. Established tools for this purpose are SnpEff (Cingolani et al. 2012) and ANNOVAR (Wang et al. 2010). 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 (Baasner et al. 2019) 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.

### 19.3.4 Mapping by Sequencing

**Forward genetics** Forward genetics describes the genetic screening of mutants which have been isolated based on an outstanding phenotype (Schneeberger 2014). 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 (Garcia et al. 2016). 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 (Schneeberger 2014).

**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) (Sikora et al. 2011). 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 (Sikora et al. 2011). 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 (Schneeberger 2014). 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 (Schneeberger et al. 2009; Cuperus et al. 2010). 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 (Wenger et al. 2010; Leshchiner et al. 2012). 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 (Abe et al. 2012). 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 (Lander and Botstein 1987). 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 (Lander and Botstein 1987; Singh et al. 2013) and no prior knowledge about the parental alleles (Voz et al. 2012) or crossing history is needed (Bowen et al. 2012). Sequencing of individual mutant genomes (Schneeberger 2014) 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 (Greenberg et al. 2011). Therefore, the resolution of MBS depends on the sampling size of correctly phenotyped and genotyped individuals in the mapping population (Schneeberger 2014). However, the resolution is only slightly affected by the number of backcrossed generations (James et al. 2013). 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 (Schneeberger and Weigel 2011). 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* (Schneeberger 2014; Schneeberger et al. 2009). Following projects applied MBS to various crop species including sugar beet (Ries et al. 2016), rice (Abe et al. 2012), maize (Liu et al. 2012), barley (Mascher et al. 2014), and cotton (Chen et al. 2015). Liu et al. applied a modification of MBS to maize for the identification of a drought tolerance locus: BSR-Seq (Liu et al. 2012). 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.

## 19.4 Transcriptomics

### 19.4.1 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 (Garalde et al. 2018). In addition, PacBio provides Iso-Seq to reveal the sequence of full length transcripts, which can facilitate gene annotation in plants (Minoche et al. 2015).

**Gene expression analysis** Short RNA-Seq reads replaced previous methods for systematic gene expression analyses like microarrays almost completely (Wang et al. 2009; Nagalakshmi et al. 2008; Mortazavi et al. 2008). Without any prior knowledge about the sequence, the abundance of transcripts can be quantified (Wang et al. 2009; Cheng et al. 2017a, b), e.g. by generating a de novo transcriptome assembly based on the RNA-Seq reads (see below) (Haak et al. 2018; Müller et al. 2017). RNA-Seq even allows to distinguish between different transcript isoforms of the same gene (Wang et al. 2009; Cheng et al. 2017a, b). Saturation of the signal as observed for microarrays is no longer an issue as the number of reads is proportional to the transcript abundance (Wang et al. 2009; Mortazavi et al. 2008). 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 (Wang et al. 2009; Hayashi et al. 2018). Transcript quantification can be performed based on alignments against a reference sequence, e.g. using STAR (Dobin et al. 2013), or alignment-free, e.g. via Kallisto (Bray et al. 2016) (Table 19.6). Information about the transcript abundance can be subjected to downstream analysis like the identification of differentially expressed genes between samples e.g. via DESeq2 (Love et al. 2014). An alternative approach is the identification of co-expressed genes or the construction of co-expression networks as described in (van Dam et al. 2017) and references therein.

**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 (Schliesky et al. 2012). De novo transcriptome assemblies were frequently applied to discover candidate genes which are responsible for a certain trait of interest (Han et al. 2017, 2018; Wu et al. 2017). One of the most popular transcriptome assemblers is Trinity (Haas et al. 2013) 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 (Hölzer and Marz 2019;



**Table 19.6** RNA-Seq gene expression tools

Name	Availability	Link	Function	References
featureCounts	Binary available	<a href="http://bioinf.wehi.edu.au/featureCounts/">http://bioinf.wehi.edu.au/featureCounts/</a>	Read counting	Liao et al. (2014)
HTSeq	Installation required	<a href="https://htseq.readthedocs.io/en/release_0.11.1/">https://htseq.readthedocs.io/en/release_0.11.1/</a>	Read counting	Anders et al. (2015)
Kallisto	Installation required	<a href="https://pachterlab.github.io/kallisto/about">https://pachterlab.github.io/kallisto/about</a>		Bray et al. (2016)
DESeq 2	R package	<a href="https://www.bioconductor.org/packages//2.12/bioc/html/DESeq2.html">https://www.bioconductor.org/packages//2.12/bioc/html/DESeq2.html</a>	Differential gene expression analysis	Love et al. (2014)
Limma	R package	<a href="https://bioconductor.org/packages/release/bioc/html/limma.html">https://bioconductor.org/packages/release/bioc/html/limma.html</a>	Differential gene expression analysis	Ritchie et al. (2015)
PIANO	R package	<a href="https://bioconductor.org/packages/release/bioc/html/piano.html">https://bioconductor.org/packages/release/bioc/html/piano.html</a>	GO/pathway enrichment analysis	Våremo et al. (2013)
WEGO	Online	<a href="http://wego.genomics.org.cn/">http://wego.genomics.org.cn/</a>	GO enrichment analysis	Ye et al. (2018)
gProfiler	Online	<a href="https://biit.cs.ut.ee/gprofiler/gost">https://biit.cs.ut.ee/gprofiler/gost</a>	GO enrichment analysis	Reimand et al. (2007)
Mercator	Online	<a href="https://www.plabipd.de/portal/web/guest/mercator4">https://www.plabipd.de/portal/web/guest/mercator4</a>	Pathway analysis	Schwacke et al. (2019)
MapMan	Online	<a href="https://plabipd.de/portal/mapman">https://plabipd.de/portal/mapman</a>	Pathway analysis	Schwacke et al. (2019)
BioMart	Online	<a href="http://plants.ensembl.org/biomart/martview">http://plants.ensembl.org/biomart/martview</a>	Pathway analysis	Smedley et al. (2015)
Plant Reactome	Online	<a href="https://plantreactome.gramene.org">https://plantreactome.gramene.org</a>	Pathway analysis	Naithani et al. (2017)

This table is an incomplete 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

Behera et al. 2017), but there are more tools that can be evaluated on a given data set (Table 19.7). Several transcriptome assemblers including Cufflinks (Trapnell et al. 2010), Trinity (Haas et al. 2013), and StringTie (Pertea et al. 2015) 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 (Wheeler et al. 2003), OrfPredictor (Min et al. 2005), and Transdecoder (Haas et al. 2013) can perform this task. Collapsing very similar sequences is sometimes required and can be performed by CD-HIT (Li and Godzik 2006; Fu et al. 2012). 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 (Bairoch and Apweiler 2000; The

**Table 19.7** De novo transcriptome assembly tools

Name	Availability	Link	References
Trinity	Installation required	<a href="https://github.com/trinityrnaseq/trinityrnaseq">https://github.com/trinityrnaseq/trinityrnaseq</a>	Haas et al. (2013)
rnaSPAdes	Binary available	<a href="http://cab.spbu.ru/software/mspades">http://cab.spbu.ru/software/mspades</a>	Bushmanova et al. (2018)
SPAdes	Binary available	<a href="http://cab.spbu.ru/software/spades">http://cab.spbu.ru/software/spades</a>	Bankevich et al. (2012)
Trans-ABYSS	Installation required	<a href="https://github.com/bcgsc/transabyss">https://github.com/bcgsc/transabyss</a>	Robertson et al. (2010)
Bridger	Installation required	<a href="https://github.com/fmaguire/Bridger_Assembler">https://github.com/fmaguire/Bridger_Assembler</a>	Chang et al. (2015)
SOAPdenovo-Trans	Installation required	<a href="https://github.com/aquaskyline/SOAPdenovo-Trans">https://github.com/aquaskyline/SOAPdenovo-Trans</a>	Xie et al. (2014)
Oases	Installation required	<a href="https://github.com/dzerbino/oases">https://github.com/dzerbino/oases</a>	Schulz et al. (2012)
IDBA-Tran	Installation required	<a href="https://github.com/loneknightpy/idba">https://github.com/loneknightpy/idba</a>	Peng et al. (2013)
BinPacker	Installation required	<a href="https://github.com/macmanes-lab/BinPacker">https://github.com/macmanes-lab/BinPacker</a>	Liu et al. (2016)
Shannon	Installation required	<a href="https://github.com/sreeramkannan/Shannon">https://github.com/sreeramkannan/Shannon</a>	Kannan et al. (2016)

This table is an incomplete list of tools which can be applied to generate plant transcriptome assemblies based on RNA-Seq data

UniProt Consortium 2017) can be harnessed to transfer the functional annotation to the newly assembled sequences. InterProScan5 (Finn et al. 2017) assigns functional annotations including gene ontology (GO) terms and identifies Pfam domains.

**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 (Dobin et al. 2013) and HISAT2 (Kim et al. 2015) are suitable tools for the mapping of RNA-Seq reads. If reads are already assembled into contigs, exonerate (Slater and Birney 2005) 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 (Slater and Birney 2005) and BLAT (Kent 2002).

## 19.5 Future Directions

Recent developments in sequencing technologies enabled the cost-efficient generation of genome and transcriptome sequences for numerous plant species of interest (Bolger et al. 2014a; Jiao and Schneeberger 2017). 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 (Jiao and Schneeberger 2017).

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 (Pucker and Brockington 2018).

Availability of plant genome sequences can foster the research on and usage of orphan crops (Chang et al. 2018) and help during de novo domestication of crops (Fernie and Yan 2019). 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

- Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H et al (2012) Genome sequencing reveals agronomically important loci in rice using MutMap. *Nat Biotechnol* 30:174–178. <https://doi.org/10.1038/nbt.2095>
- Alonso-Blanco C, Andrade J, Becker C, Bemm F, Bergelson J, Borgwardt KM et al (2016) 1135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell* 166:481–491. <https://doi.org/10.1016/j.cell.2016.05.063>
- Andrews S (2010) FastQC a quality control tool for high throughput sequence data [internet]. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed on 14 Dec 2017
- Anders S, Pyl PT, Huber W (2015) HTSeq—a python framework to work with high-throughput sequencing data. *Bioinformatics* 31:166–169. <https://doi.org/10.1093/bioinformatics/btu638>
- Baasner J-S, Howard D, Pucker B (2019) Influence of neighboring small sequence variants on functional impact prediction. <https://doi.org/10.1101/596718>
- Bairoch A, Apweiler R (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res* 28:45–48
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS et al (2012) SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol* 19:455–477. <https://doi.org/10.1089/cmb.2012.0021>
- Bayer PE, Edwards D, Batley J (2018) Bias in resistance gene prediction due to repeat masking. *Nat Plants* 4:762. <https://doi.org/10.1038/s41477-018-0264-0>
- Behera S, Voshall A, Deogun JS, Moriyama EN (2017) Performance comparison and an ensemble approach of transcriptome assembly. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 2226–2228. <https://doi.org/10.1109/bibm.2017.8218005>
- Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. *Brief Bioinform* 8:382–392. <https://doi.org/10.1093/bib/bbm048>
- Boetzer M, Pirovano W (2012) Toward almost closed genomes with GapFiller. *Genome Biol* 13:R56. <https://doi.org/10.1186/gb-2012-13-6-r56>
- Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* 27:578–579. <https://doi.org/10.1093/bioinformatics/btq683>
- Boisvert S, Laviolette F, Corbeil J (2010) Ray: Simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comput Biol* 17:1519–1533. <https://doi.org/10.1089/cmb.2009.0238>
- Bolger ME, Weisshaar B, Scholz U, Stein N, Usadel B, Mayer KF (2014a) Plant genome sequencing—applications for crop improvement. *Curr Opin Biotechnol* 26:31–37. <https://doi.org/10.1016/j.copbio.2013.08.019>
- Bolger AM, Lohse M, Usadel B (2014b) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinforma Oxford Engl.* 30:2114–2120. <https://doi.org/10.1093/bioinformatics/btu170>
- Borodovsky M, Lomsadze A (2011) Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al*. CHAPTER: Unit-4.610. <https://doi.org/10.1002/0471250953.bi0406s35>
- Bowen ME, Henke K, Siegfried KR, Warman ML, Harris MP (2012) Efficient mapping and cloning of mutations in Zebrafish by low-coverage whole-genome sequencing. *Genetics* 190:1017–1024. <https://doi.org/10.1534/genetics.111.136069>
- Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I et al (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2:10. <https://doi.org/10.1186/2047-217X-2-10>
- Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. *Nat Biotechnol* 34:525–527. <https://doi.org/10.1038/nbt.3519>
- Broad Institute (2019) Picard toolkit [internet]. Broad Institute. <https://github.com/broadinstitute/picard>

- Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol* 31:1119–1125. <https://doi.org/10.1038/nbt.2727>
- Bushmanova E, Antipov D, Lapidus A, Przhibelskiy AD (2018) rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. <https://doi.org/10.1101/420208>
- Bushnell B (2019) BBMap—Browse files at SourceForge.net. <https://sourceforge.net/projects/bbmap/files/>. Accessed on 28 May 2019
- Chaisson MJ, Tesler G (2012) Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. *BMC Bioinform* 13:238. <https://doi.org/10.1186/1471-2105-13-238>
- Chaisson MJ, Brinza D, Pevzner PA (2009) De novo fragment assembly with short mate-paired reads: does the read length matter? *Genome Res* 19:336–346. <https://doi.org/10.1101/gr.079053.108>
- Chan PP, Lin BY, Mak AJ, Lowe TM (2019) tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. <https://doi.org/10.1101/614032>
- Chang Z, Li G, Liu J, Zhang Y, Ashby C, Liu D et al (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biol* 16. <https://doi.org/10.1186/s13059-015-0596-2>
- Chang Y, Liu H, Liu M, Liao X, Sahu SK, Fu Y et al (2018) The draft genomes of five agriculturally important African orphan crops. *GigaScience*. <https://doi.org/10.1093/gigascience/giy152>
- Chen W, Yao J, Chu L, Yuan Z, Li Y, Zhang Y (2015) Genetic mapping of the nulliplex-branch gene (*gb\_nb1*) in cotton using next-generation sequencing. *Theor Appl Genet* 128:539–547. <https://doi.org/10.1007/s00122-014-2452-2>
- Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z et al (2018) The sequenced angiosperm genomes and genome databases. *Front Plant Sci* 9. <https://doi.org/10.3389/fpls.2018.00418>
- Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD (2017a) Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *Plant J* 89:789–804. <https://doi.org/10.1111/tpj.13415>
- Cheng B, Furtado A, Henry RJ (2017b) Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. *GigaScience* 6. <https://doi.org/10.1093/gigascience/gix086>
- Chikhi R, Medvedev P (2014) Informed and automated k-mer size selection for genome assembly. *Bioinformatics* 30:31–37. <https://doi.org/10.1093/bioinformatics/btt310>
- Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A et al (2016) Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* 13:1050–1054. <https://doi.org/10.1038/nmeth.4035>
- Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L et al (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* 6:80–92. <https://doi.org/10.4161/fly.19695>
- Copetti D, Búrquez A, Bustamante E, Charboneau JLM, Childs KL, Eguiarte LE et al (2017) Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proc Natl Acad Sci* 114:12003–12008. <https://doi.org/10.1073/pnas.1706367114>
- Cuperus JT, Montgomery TA, Fahlgren N, Burke RT, Townsend T, Sullivan CM et al (2010) Identification of MIR390a precursor processing-defective mutants in *Arabidopsis* by direct genome sequencing. *Proc Natl Acad Sci U S A* 107:466–471. <https://doi.org/10.1073/pnas.0913203107>
- der Auwera GAV, Carneiro MO, Hartl C, Poplin R, Angel G del, Levy-Moonshine A et al (2013) From FastQ data to high-confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinforma*. 43:11.10.1–11.10.33. <https://doi.org/10.1002/0471250953.bi1110s43>
- Dobin A, Gingeras TR (2015) Mapping RNA-seq reads with STAR. *Curr Protoc Bioinforma* 51:11.14.1–11.14.19. <https://doi.org/10.1002/0471250953.bi1114s51>
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S et al (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinforma Oxf Engl* 29:15–21. <https://doi.org/10.1093/bioinformatics/bts635>

- Dohmen E, Kremer LPM, Bornberg-Bauer E, Kemena C (2016) DOGMA: domain-based transcriptome and proteome quality assessment. *Bioinformatics* 32:2577–2581. <https://doi.org/10.1093/bioinformatics/btw231>
- El Baidouri M, Kim KD, Abernathy B, Arikrit S, Maumus F, Panaud O et al (2015) A new approach for annotation of transposable elements using small RNA mapping. *Nucleic Acids Res* 43:e84–e84. <https://doi.org/10.1093/nar/gkv257>
- Estill JC, Bennetzen JL (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes. *Plant Methods* 5:8. <https://doi.org/10.1186/1746-4811-5-8>
- Ewels P, Magnusson M, Lundin S, Käller M (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* 32:3047–3048. <https://doi.org/10.1093/bioinformatics/btw354>
- Fernie AR, Yan J (2019) De Novo domestication: an alternative route toward new crops for the future. *Mol Plant* 12:615–631. <https://doi.org/10.1016/j.molp.2019.03.016>
- Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. *Genome Res* 7:861–878. <https://doi.org/10.1101/gr.7.9.861>
- Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ et al (2017) InterPro in 2017—beyond protein family and domain annotations. *Nucleic Acids Res* 45:D190–D199. <https://doi.org/10.1093/nar/gkw1107>
- Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in De Novo annotation approaches. *PLoS ONE* 6. <https://doi.org/10.1371/journal.pone.0016526>
- Fu L, Niu B, Zhu Z, Wu S, Li W (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. <https://doi.org/10.1093/bioinformatics/bts565>
- Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio RD et al (2016) The *Cardamine hirsuta* genome offers insight into the evolution of morphological diversity. *Nat Plants* 2:16167. <https://doi.org/10.1038/nplants.2016.167>
- Galalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M et al (2018) Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* 15:201–206. <https://doi.org/10.1038/nmeth.4577>
- Garcia V, Bres C, Just D, Fernandez L, Tai FWJ, Mauxion J-P et al (2016) Rapid identification of causal mutations in tomato EMS populations via mapping-by-sequencing. *Nat Protoc* 11:2401–2418. <https://doi.org/10.1038/nprot.2016.143>
- Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. *Q-Bio*. <http://arxiv.org/abs/1207.3907>
- Glenn TC (2011) Field guide to next-generation DNA sequencers. *Mol Ecol Resour* 11:759–769. <https://doi.org/10.1111/j.1755-0998.2011.03024.x>
- Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ et al (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci* 108:1513–1518. <https://doi.org/10.1073/pnas.1017351108>
- Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 17:333–351. <https://doi.org/10.1038/nrg.2016.49>
- Greenberg MV, Ausin I, Chan SW, Cokus SJ, Cuperus JT, Feng S et al (2011) Identification of genes required for de novo DNA methylation in Arabidopsis. *Epigenetics* 6:344–354. <https://doi.org/10.4161/epi.6.3.14242>
- Haak M, Vinke S, Keller W, Droste J, Rückert C, Kalinowski J et al (2018) High quality de Novo transcriptome assembly of *Croton tiglium*. *Front Mol Biosci* 5. <https://doi.org/10.3389/fmolb.2018.00062>
- Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA et al (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol* 3(research0029):1. <https://doi.org/10.1186/gb-2002-3-6-research0029>
- Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8:1494–1512. <https://doi.org/10.1038/nprot.2013.084>

- Han Y, Wan H, Cheng T, Wang J, Yang W, Pan H et al (2017) Comparative RNA-seq analysis of transcriptome dynamics during petal development in *Rosa chinensis*. *Sci Rep* 7:43382. <https://doi.org/10.1038/srep43382>
- Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I (2018) Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. *Nat Commun* 9:619. <https://doi.org/10.1038/s41467-018-02866-0>
- Heller D, Vingron M (2018) SVM: structural variant Identification using mapped long reads. <https://doi.org/10.1101/494096>
- Hoen DR, Hickey G, Bourque G, Casacuberta J, Cordaux R, Feschotte C et al (2015) A call for benchmarking transposable element annotation methods. *Mob DNA* 6. <https://doi.org/10.1186/s13100-015-0044-6>
- Hoff KJ, Stanke M (2019) Predicting genes in single genomes with AUGUSTUS. *Curr Protoc Bioinforma* 65:e57. <https://doi.org/10.1002/cpbi.57>
- Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M (2016) BRAKER1: unsupervised RNA-seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* 32:767–769. <https://doi.org/10.1093/bioinformatics/btv661>
- Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J et al (2009) Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol* 5:e1000502. <https://doi.org/10.1371/journal.pcbi.1000502>
- Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinform* 12:491. <https://doi.org/10.1186/1471-2105-12-491>
- Hölzer M, Marz M (2019) De Novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. *GigaScience* 8. <https://doi.org/10.1093/gigascience/giz039>
- Hu Y, Green GS, Milgate AW, Stone EA, Rathjen JP, Schwessinger B (2019) Pathogen detection and microbiome analysis of infected wheat using a portable DNA sequencer. *Phytobiomes J*. <https://doi.org/10.1094/pbiomes-01-19-0004-r>
- Hunt M, Kikuchi T, Sanders M, Newbold C, Berriman M, Otto TD (2013) REAPR: a universal tool for genome assembly evaluation. *Genome Biol* 14:R47. <https://doi.org/10.1186/gb-2013-14-5-r47>
- Jain M, Fiddes I, Miga KH, Olsen HE, Paten B, Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. *Nat Methods* 12:351–356. <https://doi.org/10.1038/nmeth.3290>
- Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA et al (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 36:338–345. <https://doi.org/10.1038/nbt.4060>
- James GV, Patel V, Nordström KJ, Klasen JR, Salomé PA, Weigel D et al (2013) User guide for mapping-by-sequencing in *Arabidopsis*. *Genome Biol* 14:R61. <https://doi.org/10.1186/gb-2013-14-6-r61>
- Jiao W-B, Schneeberger K (2017) The impact of third generation genomic technologies on plant genome assembly. *Curr Opin Plant Biol* 36:64–70. <https://doi.org/10.1016/j.pbi.2017.02.002>
- Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E et al (2017) Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. *Genome Res* <https://doi.org/10.1101/gr.213652.116>
- Kajitani R, Toshimoto K, Noguchi H, Toyoda A, Ogura Y, Okuno M et al (2014) Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res* 24:1384–1395. <https://doi.org/10.1101/gr.170720.113>
- Kannan S, Hui J, Mazooji K, Pachter L, Tse D (2016) Shannon: an information-optimal de Novo RNA-Seq assembler. <https://doi.org/10.1101/039230>
- Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F (2016) Using intron position conservation for homology-based gene prediction. *Nucleic Acids Res* 44:e89. <https://doi.org/10.1093/nar/gkw092>



- Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J (2018) Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinform* 19:189. <https://doi.org/10.1186/s12859-018-2203-5>
- Kemena C, Dohmen E, Bornberg-Bauer E (2019) DOGMA: a web server for proteome and transcriptome quality assessment. *Nucleic Acids Res*. <https://doi.org/10.1093/nar/gkz366>
- Kent WJ (2002) BLAT—The BLAST-like alignment tool. *Genome Res* 12:656–664. <https://doi.org/10.1101/gr.229202>
- Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12:357–360. <https://doi.org/10.1038/nmeth.3317>
- Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L et al (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res* 22:568–576. <https://doi.org/10.1101/gr.129684.111>
- Kolmogorov M, Yuan J, Lin Y, Pevzner PA (2019) Assembly of long, error-prone reads using repeat graphs. *Nat Biotechnol* 37:540. <https://doi.org/10.1038/s41587-019-0072-8>
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*. <https://doi.org/10.1101/gr.215087.116>
- Korf I (2004) Gene finding in novel genomes. *BMC Bioinform* 5:59. <https://doi.org/10.1186/1471-2105-5-59>
- Krivtseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA et al (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* 47:D807–D811. <https://doi.org/10.1093/nar/gky1053>
- Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R et al (2016) VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res* 44:e108. <https://doi.org/10.1093/nar/gkw227>
- Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* 236:1567–1570
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. <https://doi.org/10.1038/nmeth.1923>
- Lee W-P, Stromberg MP, Ward A, Stewart C, Garrison EP, Marth GT (2014) MOSAIK: a hash-based algorithm for accurate next-generation sequencing short-read mapping. *PLoS ONE* 9:e90581. <https://doi.org/10.1371/journal.pone.0090581>
- Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* 104:520–533. <https://doi.org/10.1038/hdy.2009.165>
- Leshchiner I, Alexa K, Kelsey P, Adzhubei I, Austin-Tse CA, Cooney JD et al (2012) Mutation mapping and identification by whole-genome sequencing. *Genome Res* 22:1541–1548. <https://doi.org/10.1101/gr.135541.111>
- Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. <http://arxiv.org/abs/1303.3997>
- Li H (2016) Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinforma Oxf Engl* 32:2103–2110. <https://doi.org/10.1093/bioinformatics/btw152>
- Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinforma Oxf Engl* 34:3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li H (2019) MAQ [internet]. Available: <http://maq.sourceforge.net/maq-manpage.shtml#12>. Accessed on 28 May 2019
- Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659. <https://doi.org/10.1093/bioinformatics/btl158>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009a) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>



- Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K et al (2009b) SNP detection for massively parallel whole-genome resequencing. *Genome Res* 19:1124–1132. <https://doi.org/10.1101/gr.088013.108>
- Li C, Lin F, An D, Wang W, Huang R (2017) Genome sequencing and assembly by long reads in plants. *Genes* 9. <https://doi.org/10.3390/genes9010006>
- Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinforma Oxf Engl* 30:923–930. <https://doi.org/10.1093/bioinformatics/btt656>
- Lightfoot DJ, Jarvis DE, Ramaraj T, Lee R, Jellen EN, Maughan PJ (2017) Single-molecule sequencing and Hi-C-based proximity-guided assembly of amaranth (*Amaranthus hypochondriacus*) chromosomes provide insights into genome evolution. *BMC Biol* 15:74. <https://doi.org/10.1186/s12915-017-0412-4>
- Lin HC, Goldstein S, Mendelowitz L, Zhou S, Wetzel J, Schwartz DC et al (2012) AGORA: assembly guided by optical restriction alignment. *BMC Bioinform* 13:189. <https://doi.org/10.1186/1471-2105-13-189>
- Liu S, Yeh C-T, Tang HM, Nettleton D, Schnable PS (2012) Gene mapping via bulked segregant RNA-Seq (BSR-Seq). *PLoS ONE* 7. <https://doi.org/10.1371/journal.pone.0036406>
- Liu J, Li G, Chang Z, Yu T, Liu B, McMullen R et al (2016) BinPacker: packing-based De Novo transcriptome assembly from RNA-seq data. *PLoS Comput Biol* 12. <https://doi.org/10.1371/journal.pcbi.1004772>
- Loman NJ, Quick J, Simpson JT (2015) A complete bacterial genome assembled de novo using only nanopore sequencing data. *Nat Methods* 12:733–735. <https://doi.org/10.1038/nmeth.3444>
- Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M (2005) Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic Acids Res* 33:6494–6506. <https://doi.org/10.1093/nar/gki937>
- Lomsadze A, Burns PD, Borodovsky M (2014) Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res* 42:e119. <https://doi.org/10.1093/nar/gku557>
- Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* 15. <https://doi.org/10.1186/s13059-014-0550-8>
- Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J et al (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience* 1:18. <https://doi.org/10.1186/2047-217X-1-18>
- Mapleson D, Garcia Accinelli G, Kettleborough G, Wright J, Clavijo BJ (2017) KAT: a K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics* 33:574–576. <https://doi.org/10.1093/bioinformatics/btw663>
- Marco-Sola S, Sammeth M, Guigó R, Ribeca P (2012) The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat Methods* 9:1185–1188. <https://doi.org/10.1038/nmeth.2221>
- Mardis ER (2008a) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9:387–402. <https://doi.org/10.1146/annurev.genom.9.081307.164359>
- Mardis ER (2008b) The impact of next-generation sequencing technology on genetics. *Trends Genet TIG*. 24:133–141. <https://doi.org/10.1016/j.tig.2007.12.007>
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA et al (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376–380. <https://doi.org/10.1038/nature03959>
- Mascher M, Jost M, Kuon J-E, Himmelbach A, Abfal A, Beier S et al (2014) Mapping-by-sequencing accelerates forward genetics in barley. *Genome Biol* 15:R78. <https://doi.org/10.1186/gb-2014-15-6-r78>
- McCoy RC, Taylor RW, Blauwkamp TA, Kelley JL, Kertesz M, Pushkarev D et al (2014) Illumina TruSeq synthetic long-reads empower De Novo assembly and resolve complex, highly-repetitive transposable elements. *PLoS ONE* 9:e106689. <https://doi.org/10.1371/journal.pone.0106689>

- McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytzky A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20:1297–1303. <https://doi.org/10.1101/gr.107524.110>
- Metzker ML (2009) Sequencing in real time. *Nat Biotechnol* 27:150–151. <https://doi.org/10.1038/nbt0209-150>
- Metzker ML (2010) Sequencing technologies—the next generation. *Nat Rev Genet* 11:31–46. <https://doi.org/10.1038/nrg2626>
- Michael TP (2014) Plant genome size variation: bloating and purging DNA. *Brief Funct Genomics* 13:308–317. <https://doi.org/10.1093/bfpg/elu005>
- Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C et al (2018) High contiguity *Ara-bidopsis thaliana* genome assembly with a single nanopore flow cell. *Nat Commun* 9:541. <https://doi.org/10.1038/s41467-018-03016-2>
- Min XJ, Butler G, Storms R, Tsang A (2005) OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res* 33:W677–W680. <https://doi.org/10.1093/nar/gki394>
- Minoche AE, Dohm JC, Schneider J, Holtgräwe D, Viehöver P, Montfort M et al (2015) Exploiting single-molecule transcript sequencing for eukaryotic gene prediction. *Genome Biol* 16:184. <https://doi.org/10.1186/s13059-015-0729-7>
- Monat C, Padmarasu S, Lux T, Wicker T, Gundlach H, Himmelbach A et al (2019) TRITEX: chromosome-scale sequence assembly of Triticeae genomes with open-source tools. <https://doi.org/10.1101/631648>
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5:621–628. <https://doi.org/10.1038/nmeth.1226>
- Müller M, Seifert S, Lübke T, Leuschner C, Finkeldey R (2017) De novo transcriptome assembly and analysis of differential gene expression in response to drought in European beech. *PLoS ONE* 12:e0184167. <https://doi.org/10.1371/journal.pone.0184167>
- Myers JEW (2016) A history of DNA sequence assembly. *It-Inf Technol* 58:126–132. <https://doi.org/10.1515/itit-2015-0047>
- Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M et al (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science* 320:1344–1349. <https://doi.org/10.1126/science.1158441>
- Naithani S, Preece J, D'Eustachio P, Gupta P, Amarasinghe V, Dharmawardhana PD et al (2017) Plant Reactome: a resource for plant pathways and comparative analysis. *Nucleic Acids Res* 45:D1029–D1039. <https://doi.org/10.1093/nar/gkw932>
- Nawrocki EP, Eddy SR (2013) Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29:2933–2935. <https://doi.org/10.1093/bioinformatics/btt509>
- Nguyen SH, Duarte TPS, Coin LJM, Cao MD (2017) Real-time demultiplexing Nanopore bar-coded sequencing data with npBarcode. *Bioinformatics* 33:3988–3990. <https://doi.org/10.1093/bioinformatics/btx537>
- NovoCraft (2010) NovoAlign [internet]. <http://www.novocraft.com/products/novoalign/>. Accessed on 27 May 2019
- Paulino D, Warren RL, Vandervalk BP, Raymond A, Jackman SD, Birol I (2015) Sealer: a scalable gap-closing application for finishing draft genomes. *BMC Bioinform* 16. <https://doi.org/10.1186/s12859-015-0663-4>
- Payne A, Holmes N, Rakyan V, Loose M (2018) Whale watching with BulkVis: a graphical viewer for Oxford Nanopore bulk fast5 files. <https://doi.org/10.1101/312256>
- Peng Y, Leung HCM, Yiu S-M, Lv M-J, Zhu X-G, Chin FYL (2013) IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics* 29:i326–i334. <https://doi.org/10.1093/bioinformatics/btt219>
- Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* 33:290–295. <https://doi.org/10.1038/nbt.3122>

- Peterson DG, Arick M (2018) Sequencing plant genomes, pp 1–85. [https://doi.org/10.1007/124\\_2018\\_18](https://doi.org/10.1007/124_2018_18)
- Phillippy AM (2017) New advances in sequence assembly. *Genome Res* 27:xi–xiii. <https://doi.org/10.1101/gr.223057.117>
- Pomerantz A, Peñafiel N, Arteaga A, Bustamante L, Pichardo F, Coloma LA et al (2018) Real-time DNA barcoding in a rainforest using nanopore sequencing: opportunities for rapid biodiversity assessments and local capacity building. *GigaScience* 7. <https://doi.org/10.1093/gigascience/gy033>
- Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A et al (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat Biotechnol* 36:983–987. <https://doi.org/10.1038/nbt.4235>
- Price AL, Jones NC, Pevzner PA (2005) De novo identification of repeat families in large genomes. *Bioinforma Oxf Engl* 21(Suppl 1):i351–i358. <https://doi.org/10.1093/bioinformatics/bti1018>
- Pucker B, Brockington SF (2018) Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genom* 19:980. <https://doi.org/10.1186/s12864-018-5360-z>
- Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B (2016) A De Novo genome sequence assembly of the *Arabidopsis thaliana* accession Niederzenz-1 displays presence/absence variation and strong synteny. *PLoS ONE* 11:e0164321. <https://doi.org/10.1371/journal.pone.0164321>
- Pucker B, Holtgräwe D, Weisshaar B (2017) Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. *BMC Res Notes* 10. <https://doi.org/10.1186/s13104-017-2985-y>
- Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R et al (2019) A chromosome-level sequence assembly reveals the structure of the *Arabidopsis thaliana* Nd-1 genome and its gene set. *PLoS ONE* 14:e0216233. <https://doi.org/10.1371/journal.pone.0216233>
- QIAGEN (2016) QIAGEN bioinformatics—sample to Insight. In: QIAGEN bioinformatics [internet]. <https://www.qiagenbioinformatics.com/>. Accessed 16 Dec 2018
- Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR et al (2012) A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genom* 13:341. <https://doi.org/10.1186/1471-2164-13-341>
- Reimand J, Kull M, Peterson H, Hansen J, Vilo J (2007) g:profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* 35:W193–W200. <https://doi.org/10.1093/nar/gkm226>
- Ries D, Holtgräwe D, Viehöver P, Weisshaar B (2016) Rapid gene identification in sugar beet using deep sequencing of DNA from phenotypic pools selected from breeding panels. *BMC Genomics* 17. <https://doi.org/10.1186/s12864-016-2566-9>
- Rimmer A, Phan H, Mathieson I, Iqbal Z, Twigg SRF, Wgs500 Consortium, et al (2014) Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat Genet.* 46:912–918. <https://doi.org/10.1038/ng.3036>
- Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W et al (2015) limma powers differential expression analyses for RNA-seq and microarray studies. *Nucleic Acids Res* 43:e47. <https://doi.org/10.1093/nar/gkv007>
- Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD et al (2010) De novo assembly and analysis of RNA-seq data. *Nat Methods* 7:909–912. <https://doi.org/10.1038/nmeth.1517>
- Ruan J, Li H (2019) Fast and accurate long-read assembly with wtdbg2. <https://doi.org/10.1101/530972>
- Saha S, Bridges S, Magbanua ZV, Peterson DG (2008) Computational approaches and tools used in identification of dispersed repetitive DNA sequences. *Trop Plant Biol* 1:85–96. <https://doi.org/10.1007/s12042-007-9007-5>
- Salzberg SL (2019) Next-generation genome annotation: we still struggle to get it right. *Genome Biol* 20:92. <https://doi.org/10.1186/s13059-019-1715-2>

- Salzberg SL, Phillippy AM, Zimin A, Puiu D, Magoc T, Koren S et al (2012) GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res* 22:557–567. <https://doi.org/10.1101/gr.131383.111>
- Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J Mol Biol* 94:441–448
- Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74:5463–5467
- Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. *Hum Mol Genet* 19:R227–R240. <https://doi.org/10.1093/hmg/ddq416>
- Schliesky S, Gowik U, Weber APM, Bräutigam A (2012) RNA-Seq assembly—are we there yet? *Front Plant Sci* 3:220. <https://doi.org/10.3389/fpls.2012.00220>
- Schneeberger K (2014) Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat Rev Genet* 15:662–676. <https://doi.org/10.1038/nrg3745>
- Schneeberger K, Weigel D (2011) Fast-forward genetics enabled by new sequencing technologies. *Trends Plant Sci* 16:282–288. <https://doi.org/10.1016/j.tplants.2011.02.006>
- Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL et al (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat Methods* 6:550–551. <https://doi.org/10.1038/nmeth0809-550>
- Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28:1086–1092. <https://doi.org/10.1093/bioinformatics/bts094>
- Schwacke R, Ponce-Soto GY, Krause K, Bolger AM, Arsova B, Hallab A et al (2019) MapMan4: a refined protein classification and annotation framework applicable to multi-omics data analysis. *Mol Plant*. <https://doi.org/10.1016/j.molp.2019.01.003>
- Sedlazeck FJ, Rescheneder P, von Haeseler A (2013) NextGenMap: fast and accurate read mapping in highly polymorphic genomes. *Bioinformatics* 29:2790–2791. <https://doi.org/10.1093/bioinformatics/btt468>
- Shariat B, Movahedi NS, Chitsaz H, Boucher C (2014) HyDA-Vista: towards optimal guided selection of k-mer size for sequence assembly. *BMC Genom* 15:S9. <https://doi.org/10.1186/1471-2164-15-S10-S9>
- Shen Y, Wan Z, Coarfa C, Drabek R, Chen L, Ostrowski EA et al (2010) A SNP discovery method to assess variant allele probability from next-generation resequencing data. *Genome Res* 20:273–280. <https://doi.org/10.1101/gr.096388.109>
- Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* 26:1135–1145. <https://doi.org/10.1038/nbt1486>
- Shendure J, Mitra RD, Varma C, Church GM (2004) Advanced sequencing technologies: methods and goals. *Nat Rev Genet* 5:335–344. <https://doi.org/10.1038/nrg1325>
- Sikora P, Chawade A, Larsson M, Olsson J, Olsson O (2011) Mutagenesis as a tool in plant genetics, functional genomics, and breeding. *Int J Plant Genomics* <https://doi.org/10.1155/2011/314829>
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinforma Oxf Engl* 31:3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Simpson JT, Durbin R (2012) Efficient de novo assembly of large genomes using compressed data structures. *Genome Res* 22:549–556. <https://doi.org/10.1101/gr.126953.111>
- Singh R, Leslie Low E-T, Ooi LC-L, Ong-Abdullah M, Chin TN, Nagappan J et al (2013) The oil palm Shell gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature* 500:340–344. <https://doi.org/10.1038/nature12356>
- Slater GS, Birney E (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinform* 6:31. <https://doi.org/10.1186/1471-2105-6-31>
- Smedley D, Haider S, Durinck S, Pandini L, Provero P, Allen J et al (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res* 43:W589–W598. <https://doi.org/10.1093/nar/gkv350>

- Smit A, Hubley R, Green P (2015) RepeatMasker frequently open-4.0 [internet]. <http://www.repeatmasker.org/>
- Smith AD, Chung W-Y, Hodges E, Kendall J, Hannon G, Hicks J et al (2009) Updates to the RMAP short-read mapping software. *Bioinformatics* 25:2841–2842. <https://doi.org/10.1093/bioinformatics/btp533>
- Somssich M (2018) A short history of *Arabidopsis thaliana* (L.) Heynh. Columbia-0 [internet]. PeerJ Inc. Report No. e26931v4. <https://doi.org/10.7287/peerj.preprints.26931v4>
- Souvorov A, Kapustin Y, Kiryutin B, Chetvermin V, Tatusova T, Lipman D (2010) Gnomon—NCBI eukaryotic gene prediction tool. <http://www.ncbi.nlm.nih.gov/core/assets/genome/files/Gnomon-description.pdf>
- Sović I, Šikić M, Wilm A, Fenlon SN, Chen S, Nagarajan N (2016) Fast and sensitive mapping of nanopore sequencing reads with GraphMap. *Nat Commun* 7:11307. <https://doi.org/10.1038/ncomms11307>
- Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res* 34:W435–W439. <https://doi.org/10.1093/nar/gkl200>
- Stoiber M, Brown J (2017) BasecRAWller: streaming nanopore basecalling directly from raw signal. <https://doi.org/10.1101/133058>
- Szalay T, Golovchenko JA (2015) De novo sequencing and variant calling with nanopores using PoreSeq. *Nat Biotechnol* 33:1087–1091. <https://doi.org/10.1038/nbt.3360>
- Tang H, Lyons E, Town CD (2015) Optical mapping in plant comparative genomics. *GigaScience* 4. <https://doi.org/10.1186/s13742-015-0044-y>
- Ter-Hovhannisyanyan V, Lomsadze A, Chernoff YO, Borodovsky M (2008) Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res* 18:1979–1990. <https://doi.org/10.1101/gr.081612.108>
- The UniProt Consortium (2017) UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 45:D158–D169. <https://doi.org/10.1093/nar/gkw1099>
- Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, Van MB et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28:511–515. <https://doi.org/10.1038/nbt.1621>
- Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR et al (2018) Evaluation of Oxford Nanopore’s MinION sequencing device for microbial whole genome sequencing applications. *Sci Rep* 8:10931. <https://doi.org/10.1038/s41598-018-29334-5>
- van Dam S, Vösa U, van der Graaf A, Franke L, de Magalhães JP (2017) Gene co-expression analysis for functional classification and gene–disease predictions. *Brief Bioinform* 19:575–592. <https://doi.org/10.1093/bib/bbw139>
- Väremo L, Nielsen J, Nookaew I (2013) Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Res* 41:4378–4391. <https://doi.org/10.1093/nar/gkt111>
- Vaser R, Sović I, Nagarajan N, Šikić M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* 27:737–746. <https://doi.org/10.1101/gr.214270.116>
- Vicent CM, Casacuberta JM (2017) Impact of transposable elements on polyploid plant genomes. *Ann Bot* 120:195–207. <https://doi.org/10.1093/aob/mcx078>
- Voz ML, Coppiters W, Manfroid I, Baudhuin A, Von Berg V, Charlier C et al (2012) Fast homozygosity mapping and identification of a Zebrafish ENU-induced mutation by whole-genome sequencing. *PLoS ONE* 7. <https://doi.org/10.1371/journal.pone.0034671>
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S et al (2014) Pilon: an Integrated Tool for Comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 9:e112963. <https://doi.org/10.1371/journal.pone.0112963>
- Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10:57–63. <https://doi.org/10.1038/nrg2484>

- Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164. <https://doi.org/10.1093/nar/gkq603>
- Waterhouse RM, Seppey M, Simão FA, Manni M, Ioannidis P, Klioutchnikov G et al (2018) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol* 35:543–548. <https://doi.org/10.1093/molbev/msx319>
- Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. *Nucleic Acids Res* 39:e132–e132. <https://doi.org/10.1093/nar/gkr599>
- Wenger JW, Schwartz K, Sherlock G (2010) Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae*. *PLoS Genet* 6. <https://doi.org/10.1371/journal.pgen.1000942>
- Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU et al (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res* 31:28–33
- Wilm A, Aw PPK, Bertrand D, Yeo GHT, Ong SH, Wong CH et al (2012) LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res* 40:11189–11201. <https://doi.org/10.1093/nar/gks918>
- Wu S, Lei J, Chen G, Chen H, Cao B, Chen C (2017) De novo transcriptome assembly of chinese kale and global expression analysis of genes involved in glucosinolate metabolism in multiple tissues. *Front Plant Sci* 8. <https://doi.org/10.3389/fpls.2017.00092>
- Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S et al (2014) SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics* 30:1660–1666. <https://doi.org/10.1093/bioinformatics/btu077>
- Ye J, Zhang Y, Cui H, Liu J, Wu Y, Cheng Y et al (2018) WEGO 2.0: a web tool for analyzing and plotting GO annotations, update. *Nucleic Acids Res* 2018(46):W71–W75. <https://doi.org/10.1093/nar/gky400>
- Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res* 18:821–829. <https://doi.org/10.1101/gr.074492.107>
- Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. *Bioinforma Oxf Engl*. 29:2669–2677. <https://doi.org/10.1093/bioinformatics/btt476>