De novo Nd-1 genome assembly reveals genomic diversity of *Arabidopsis thaliana* and facilitates genome-wide non-canonical splice site analyses across plant species

DISSERTATION

submitted by

Boas Pucker

for the degree Doctor of Science (Dr. rer. nat.)

Faculty of Biology, Bielefeld University

January 2019

The work of this thesis was performed at the chair of Genome Research and later Genetics and Genomics of Plants (CeBiTec & Faculty of Biology, Bielefeld University) and in the Evolution and Diversity group (Department of Plant Sciences, University of Cambridge) from February 2016 until November 2018 under the supervision of Prof. Dr. Bernd Weisshaar (Genetics and Genomics of Plants, Bielefeld University) and Prof. Dr. Jörn Kalinowski (Microbial Genomics and Biotechnology, Bielefeld University).

Parts of the results of this thesis were published in:

Pucker, B., Holtgräwe, D., Rosleff Sörensen, T., Stracke, R., Viehöver, P., and 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. doi:10.1371/journal.pone.0164321.

Pucker, B., Holtgräwe, D., and Weisshaar, B. (2017). Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. BMC Research Notes, 10, 667. doi:10.1186/s13104-017-2985-y.

Pucker, B. and Brockington, S.F. (2018). Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. *BMC Genomics*. 2018;19(1). doi:10.1186/s12864-018-5360-z.

Parts of the results of this thesis are currently under review:

Pucker, B., Holtgraewe, D., Stadermann, K. B., Frey, K., Huettel, B., Reinhardt, R., and Weisshaar, B. A Chromosome-level Sequence Assembly Reveals the Structure of the *Arabidopsis thaliana* Nd-1 Genome and its Gene Set. Submitted to PLOS ONE.

Abbreviations

BAC	bacterial artificial chromosome
BLAST	basic local alignment search tool
cDNA	copy deoxyribonucleic acid
CNV	copy number variation
DBG	De Bruijn graph
DNA	deoxyribonucleic acid
ESTs	expressed sequence tags
kbp	kilo base pairs
LTR	long terminal repeat
Mbp	million base pairs
mRNA	messenger ribonucleic acid
Mya	million years ago
NOR	nucleolus organizing region
OLC	overlap layout consensus
ONT	Oxford Nanopore Technologies
PacBio	Pacific Biosciences
PAV	presence/absence variation
PCR	polymerase chain reaction
QTL	quantitative trait loci
rDNA	ribosomal DNA (encodes rRNA)
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
SMRT	single molecule real-time
SNP	single nucleotide polymorphism

SV	structural variant
T-DNA	transfer deoxyribonucleic acid
TE	transposable element
tRNA	transfer ribonucleic acid

Gene names, units, and common abbreviations are not included.

Table of Content

Abstract **1**

1 Introduction **4**

1.1 Plant genome and transcriptome sequencing 4

1.1.1 Motivation and application 4

1.1.2 Generations of sequencing technologies and strategies 6

1.1.2.1 Second generation sequencing technologies 7

1.1.2.2 Whole genome shotgun sequencing 10

1.1.2.3 Long reads of the third generation 11

1.1.3 RNA-Seq 13

1.2 Genome assembly 14

1.2.1 Assembly of reads into contigs and scaffolds 14

1.2.2 Linkage information for high-level assembly scaffolding 17

1.2.3 Genome assembly validation 18

1.3 Genome sequence annotation 19

1.3.1 Prediction of gene structures 19

1.3.2 Annotation of transposable elements 20

1.4 Re-sequencing projects and the diversity within species 21

1.5 Arabidopsis thaliana – a model organism for plant genomics 22

1.5.1 Columbia-0, Landsberg *erecta*, and Niederzenz-1 24

1.5.2 Genetics and genomics of *A. thaliana* 25

1.5.3 Molecular evolution 26

1.5.4 Genome size of *A. thaliana* 26

1.5.5 Transposable elements in *A. thaliana* 27

1.5.6 Gene set of *A. thaliana* 27

1.5.7 Transcriptomics of *A. thaliana* 29

1.6 Splicing and splice sites 29

1.7 Objectives 31

2 Results **32**

2.1 *De novo* genome sequence assembly of *A. thaliana* Nd-1 32

2.1.1 The *A. thaliana* Nd-1 assembly ... 33
2.1.2 Small sequence variants ... 35
2.1.3 Presence/absence variations ... 37
2.2 Consideration of non-canonical splice sites improves gene prediction ... 38
2.2.1 Non-canonical splice sites in *A. thaliana* ... 39
2.2.2 Improved gene prediction ... 39
2.3 Chromosome-level assembly reveals structural variants between Nd-1 and Col-0 and facilitates pan-genomic analyses ... 41
2.3.1 Assembly based on SMRT sequencing reads ... 41
2.3.2 Large structural variants ... 44
2.3.3 The Nd-1 gene set ... 45
2.3.4 Pan-genomic analysis of *A. thaliana* ... 45
2.4 Non-canonical splice sites in plant genomes ... 46
2.4.1 Annotated diversity ... 46
2.4.2 Intron sizes in relation to splice site combinations ... 47
2.4.3 Validation and usage quantification of splice sites ... 48
2.4.4 Script collection for the investigation of splice site combinations ... 49
3 Discussion and outlook ... 50
3.1 Genome sequencing and assembly ... 50
3.2 Genome size of *A. thaliana* ... 52
3.3 Gene set differences ... 55
3.4 Non-canonical splice sites ... 56
3.5 Transfer to crops ... 58
3.6 Conclusion ... 60
4 References ... 61
5 Acknowledgements ... 79
6 Supplements ... 80

Abstract

Arabidopsis thaliana is currently the most important plant model organism and therefore frequently used to investigate processes, which are more complex in other plants. The *A. thaliana* Columbia-0 (Col-0) genome sequence was the first available one of all plants [1] and comes with a high quality annotation [2]. Despite the use of numerous *A. thaliana* accessions in research projects, no other genome sequence of this species was available for a long time. Pan-genomic investigations were restricted to re-sequencing studies mainly limited by the available sequencing capacities. This hampered the discovery of large structural variants and investigations of genome evolution. Substantial technological progress during the last years made sequencing and *de novo* assembly of plant genomes feasible - even for single research groups. Since genes are determining the phenotype of a plant species, they are often the focus of genome sequencing projects. One major challenge during the prediction of protein encoding genes is the accurate detection of splice sites. Although terminal dinucleotides in introns are well conserved on the genomic level with GT at the 5'-end and AG at the 3'-end, there are a few reports about some rare variations [3,4]. Because of the extremely high number of possible gene models when considering splice site combinations besides this canonical GT-AG combination, *ab initio* gene prediction cannot identify non-canonical splice site combinations.

Objectives of this work were i) the generation of a high quality *A. thaliana* Niederzenz-1 (Nd-1) genome sequence assembly with a corresponding annotation and comparison against the Col-0 reference genome sequence, ii) investigation of non-canonical splice sites in *A. thaliana*, and iii) transfer of methods and knowledge about splice sites to the investigation of non-canonical splice sites across annotated plant genome sequences.

Abstract

The following points summarize key results of this work:

High quality *A. thaliana* Nd-1 genome sequence and corresponding annotations

- Based on single molecule real-time sequencing reads, 123.5 Mbp of the *A. thaliana* Nd-1 genome sequence were assembled with an N50 of 13.4 Mbp. Successful identification of benchmarking sequences and high mapping rates of expressed sequence tags indicate a high assembly quality.
- Hint-based gene prediction was applied to consider non-canonical splice sites in the gene prediction process and resulted in a final set of 27,247 protein encoding genes. This structural annotation is considered to be high quality as more than 89% of the nuclear protein encoding genes in the Araport11 annotation of the Col-0 reference sequence were matched as reciprocal best BLAST hits.
- Comparison of the Nd-1 and Col-0 genome sequences revealed large structural variants often in proximity to the centromeres. With approximately 1 Mbp in length an inversion in the north of chromosome 4 is currently the biggest difference seen. A collapsed region in the Col-0 genome sequence around At4g22214 was detected during validation of apparent tandem duplication differences.

Investigation of non-canonical splice sites in *A. thaliana*

- In total, 1,267 representative transcripts of the Araport11 annotation contain non-canonical splice sites. Therefore, about 5% of all nuclear protein encoding genes in Araport11 cannot be predicted accurately without the consideration of non-canonical splice sites.
- Canonical GT-AG splice site combinations are present in 98.9% of all introns. The major non-canonical splice site combinations GC-AG (0.9%) and AT-AC (0.1%) account for the biggest proportion of non-canonical splice site combinations in *A. thaliana*. Diverse minor non-canonical splice site combinations account for the remaining 0.1% of all splice site combinations.
- RNA-Seq reads and cDNA-based amplicon sequencing support minor non-canonical splice site combinations. Genes with validated non-canonical splice site combinations contain on average ten exons thus substantially exceeding the average of four exons per gene.

Investigation of non-canonical splice sites across the kingdom of plants

- The combined frequency of all minor non-canonical splice site combinations (0.09%) substantially exceeds the frequency of the major non-canonical AT-AC splice site combinations (0.06%).
- Minor non-canonical splice site combinations are not just artefacts, but supported by RNA-Seq reads in multiple plant species. Moreover, the sequences of non-canonical splice site combinations are non-random displaying a strong decrease in frequency with divergence from the canonical GT-AG splice site combination.
- Donor splice sites displayed a stronger usage compared to acceptor splice sites indicating that there might be more flexibility in the splicing process at the 3'-end of an intron.

1 Introduction

This introduction provides the general background for the analyses, results, and discussions in the following sections of the thesis. First, the motivation for sequencing plant genomes and transcriptomes as well as the progress in these fields are described. Different sequencing technologies for the analyses of DNA and RNA are presented, because this work is focused on sequence analysis of these molecule types. Descriptions of bioinformatic concepts and tools for the processing of the resulting data sets follow. Current challenges like functional annotation and diversity investigations by comprehensive re-sequencing projects are pointed out. The model organism *Arabidopsis thaliana* is introduced by describing selected aspects of previous work in genetics and genomics. Finally, an introduction into splicing and the importance of splice sites closes this section.

1.1 Plant genome and transcriptome sequencing

Besides the beautiful appearance of many flowers, plants are important due to various ecosystem services like CO_2 fixation and protection of soil against erosion. Especially the contribution to the human nutrition is economically relevant. Understanding the genetic and genomic basis of plant biology is the first step towards the improvement of traits in breeding projects.

1.1.1 Motivation and application

Genome and transcriptome sequences are powerful resources for the plant research community, as comprehensive insights into species can be inferred. Sequence-based approaches range from oligonucleotide design [5] to RNA-Seq experiments [6–9]. Genome-wide investigations of gene families [10–12] are also facilitated by available genome and transcriptome sequences. Comparative genomics enables the identification of structural variants (SVs) [13–15], the assessment of diversity in a population [16–18], the identification of genomic regions under selection [19,20], and the investigation of genome evolution [21–25]. Genome sequences are crucial for the efficient development of molecular markers e.g. for the detection of quantitative trait loci (QTL) in research [26–28], marker-assisted selection in crop breeding [29–34], or even to enable genetic engineering of plants [35,36]. Advanced breeding approaches [37–39] often rely on availability of genome sequences and suitable annotations. Even partial and fragmented genome or transcriptome sequences can be valuable when enabling the development of molecular markers to increase the resolution of genetic mapping approaches [40]. Making genomic resources available can help to establish new crop species [34,41,42]. Although these orphan crop species [43] are of minor economic and scientific interest, recent advances in sequencing technologies enable the cost-effective generation of genome or transcriptome sequences. Pan-genomic resources, i.e. multiple genome sequences of the same species, can facilitate the conservation of genetic diversity and provide economic benefits when used to advice crosses between landraces and wild relatives [44–47]. To harness the full potential of sequences, a structural and functional annotation is required. As the annotation process of new genome sequences is often based on comprehensive and reliable annotations of other plant genome sequences [48,49], the generation of high quality annotations for several model plant species is of high importance. In summary, these examples show the relevance of genome and transcriptome sequencing projects in facilitating basic research and crop improvements.

Recent publications provided numbers of sequenced plant genomes [50–52], but these are quickly outdated due to dropping sequencing costs (Fig.1). Rapid developments in sequencing technologies enable almost every research group to run own sequencing projects [51,53,54]. Therefore, it is no longer feasible to provide exact information about the number of sequenced plant genomes. Recent announcements by leading sequencing centres indicate that genomic resources for all living species might be available within a couple of years.

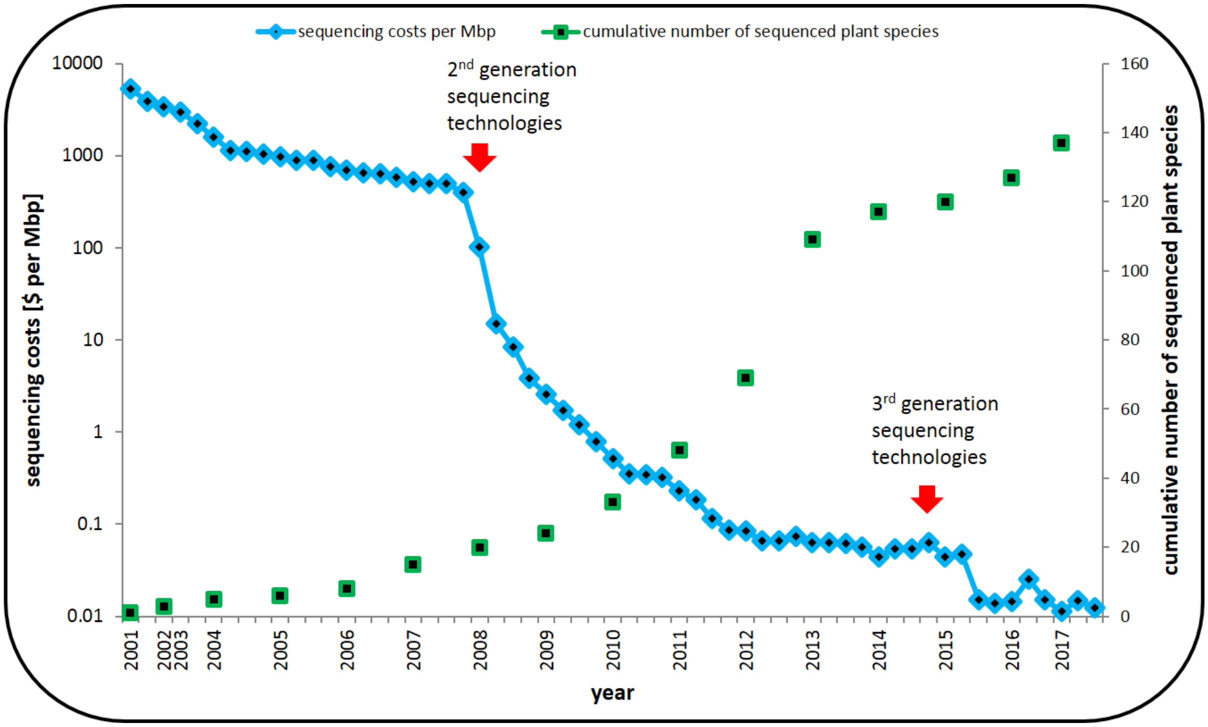

Fig.1: Sequencing costs and number of sequenced plant species.

This development of per Mbp sequencing costs from 2001 to 2017 is based on data provided by the National Human Genome Research Institute [55]. Basis of the presented values are the production costs for sequence generation without any downstream processing. With substantially dropping sequencing costs, the number of sequenced plant species increases. Since not all generated genome sequences are published and only the first complete sequence per species is counted, the presented values are lower bounds [56].

1.1.2 Generations of sequencing technologies and strategies

Sanger sequencing [57,58] and the method developed by Maxam and Gilbert [59] are usually considered as the first generation of sequencing technologies [60]. Although the chain-termination method developed by Sanger is still frequently applied e.g. for the validation of constructed plasmids or the investigation of amplicons [61–66], methods of the first generation are omitted here for brevity. Unfortunately, sequencing technologies of the following generations are inconsistently classified in the literature [60,67–74]. Throughout this work, Roche/454 pyrosequencing, Solexa/Illumina® sequencing-by-synthesis, and Ion Torrent sequencing are considered as second generation. Ion Torrent will be considered as second generation due to the short read lengths which are closer to Roche/454 reads than to long reads generated by Oxford Nanopore Technologies (ONT). Although the concept of nanopore sequencing goes back to the 1980s [75], the two long read sequencing technologies provided by Pacific Biosciences (PacBio®) and ONT will be described as third generation. Despite timely overlap between the technologies of all generations [60,76], the third generation is currently dominating genome sequencing approaches due to extremely high contiguity achieved in long read assemblies. Nevertheless, second generation technologies are still deployed in applications where cost-efficient generation of numerous tags is more important than the length of reads e.g. RNA-Seq [77].

1.1.2.1 Second generation sequencing technologies

Sequencing technologies of the second generation were dominated by Roche/454 and Solexa/Illumina technologies [78–82]. Due to the origin after the first generation of sequencing technologies, second generation sequencing technologies are frequently referred to as ‘next generation sequencing’ (NGS).

Roche/454 pyrosequencing is based on the release of pyrophosphate upon integration of a nucleotide into the synthesized DNA strand which is detected based on a chain of enzymatic reactions ultimately resulting in luminescence emitted from a firefly luciferase [60,78,82]. Reactions are taking place in extremely small wells containing only copies of one template molecule, which was previously amplified via so called emulsion PCR inside extremely small water droplets embedded in oil. The sequencing process involves i) the successive streaming of nucleotides in a predetermined order (TCGA), ii) the continuous measuring of fluorescence as result of a nucleotide incorporation, and iii) extensive washing steps to keep the background signal low [60,78]. Although luminescence signal intensity corresponds to the number of integrated nucleotides, it reaches saturation in homopolymers leading to length errors [83]. Despite this drawback, the read length of Roche/454 sequencing substantially exceeded the achievements of all other second generation sequencing technologies at that time.

The Ion Torrent sequencing is based on semiconductor technology that allows the detection of protons when nucleotides are incorporated into a DNA strand [84]. After attachment of templates to a bead, amplification is performed similar to Roche/454, and following sequencing is performed with one bead per well in a microtiter plate [72]. Nucleotides are supplied successively with washing steps in between to keep the noise low. Like Roche/454, the Ion Torrent technology is likely to produce sequencing errors in homopolymers [71], because the pH signal is only roughly proportional to the number of integrated nucleotides [72]. In addition, it is not suitable for sequencing AT-rich regions [71], which are frequent in plant genomes. The biggest advantages are the extremely short run time of only a few hours and the lack of optics, which facilitate sequencing outside the lab under less controlled conditions [72].

The Solexa/Illumina technology is sequencing by cycle reversible termination on a surface (Fig.2) [60,82]. Libraries are generated by adding adapters to DNA fragments and amplifying these in an initial PCR step. Next, these adapters bind to complementary sequences on the surface of dedicated flow cells. Bridge amplification on the flow cell is used to generate clusters of molecule copies which amplify the signal during the following sequencing steps. Sequencing is performed by supplying nucleotides marked with four specific fluorescence dyes which block the 3'-OH thus permitting only the incorporation of a single nucleotide per cycle [85,86]. After fluorescence readout, the block is removed to enable the integration of the next nucleotide [79]. The incorporation of a single nucleotide in each cycle results in a clear fluorescence signal per cluster and finally equal read lengths from all clusters. Despite this elegant design and generally low error rates, several systematic errors were identified [87]. Systematic errors include an increased error probability after 'G' [88] especially following the 'GGC' motif [71,89] and underrepresentation of regions with a very low [88,90,91] or very high GC content [90,91].

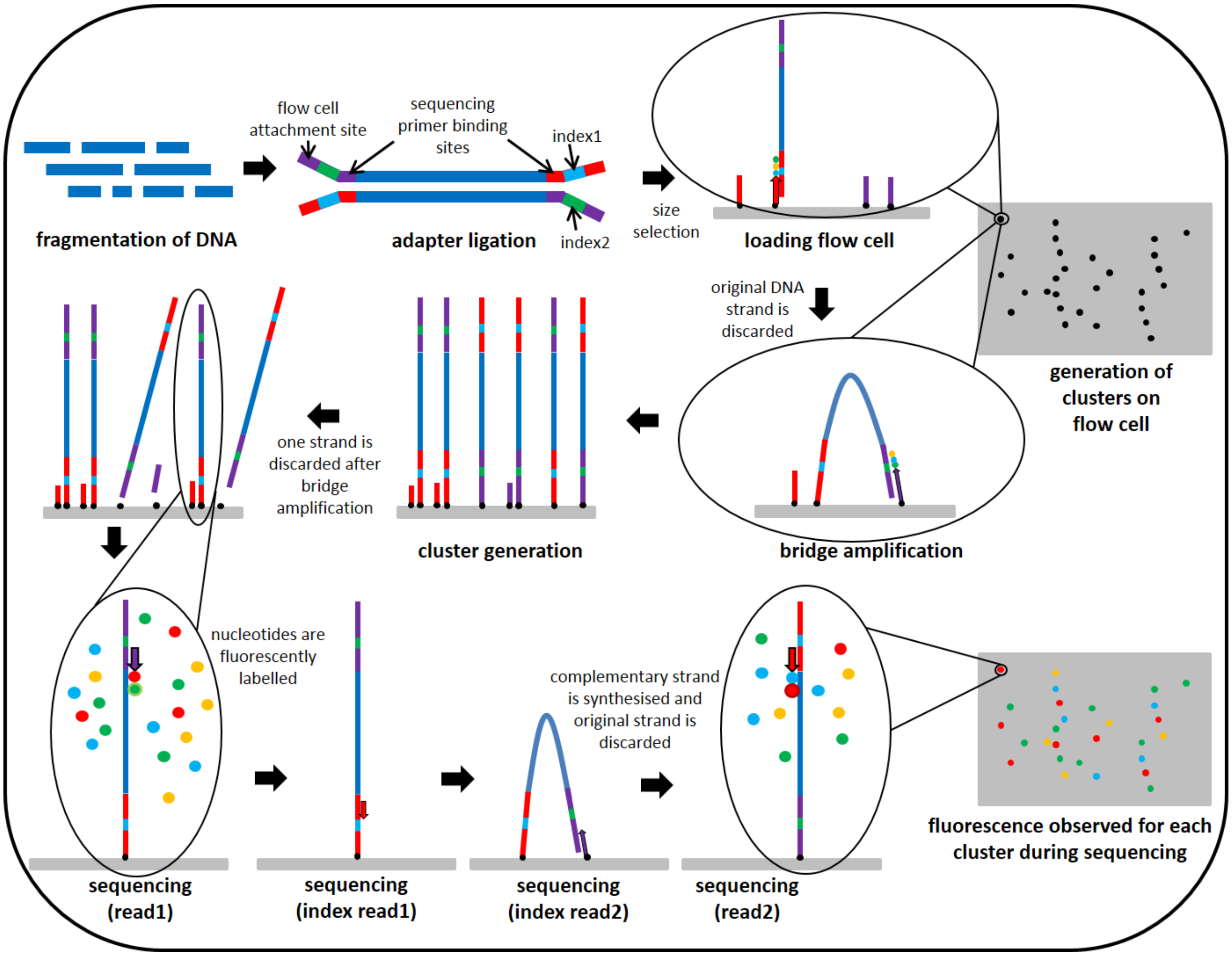
fragmentation of DNA
flow cell attachment site
sequencing primer binding sites
index1
index2
adapter ligation
size selection
loading flow cell
original DNA strand is discarded
generation of clusters on flow cell
bridge amplification
cluster generation
one strand is discarded after bridge amplification
nucleotides are fluorescently labelled
complementary strand is synthesised and original strand is discarded
fluorescence observed for each cluster during sequencing
sequencing (read1)
sequencing (index read1)
sequencing (index read2)
sequencing (read2)

Fig.2: Illumina sequencing.

Simplified illustration of an illumina paired-end sequencing process of a dual-indexed library. Several steps including washing, strand removals, and the synthesis of a complementary strand prior to the second sequencing process are not shown. Although multiple copies of identical molecules are sequenced in parallel, these steps are only displayed for one template molecule.

Illumina sequencing is still applied in genome sequencing projects as it is cost-efficient and generates reads with extremely low error rates. Sequencing of DNA fragments from both ends (paired-end) is a frequently applied mode. The distance between reads is determined by the size of the DNA fragment enclosed by adapters at both ends, i.e. the insert size. Mate pair sequencing is a more sophisticated protocol developed to generate read pairs with even larger insert sizes [60]. Several kbp long DNA fragments are tagged at the ends and circularized. Next, the circular DNA molecule is fragmented and fragments with joints of the original ends are enriched. This enriched fraction is subjected to paired-end sequencing. The resulting reads are orientated in opposite directions, but can be converted into paired-end read pairs through bioinformatic processing [92]. A methods for the investigation of the three-dimensional DNA structure i.e. Hi-C [93] involves the generation of read pairs with even larger distances although the distance of reads in a pair is only roughly known [94,95]. Hi-C relies on the assumption that DNA parts in close proximity in space are also close together on the same DNA strand [93]. Based on this assumption, chromatin is isolated, chimera DNA molecules are generated from neighbouring molecules, and cleaved by restriction enzymes. The resulting chimeric DNA fragments are subjected to paired-end sequencing.

PCR can be used to selectively amplify fragments and allows very small amounts of DNA as starting material for sequencing processes, but it is biased in several ways [91,96–98]. To minimize the biases introduced by PCR amplification during sequencing library preparation, PCR-free protocols were developed [98]. The bridge amplification of the flow cell is enriching fragments with successful ligated adapters at both ends thus avoiding an additional PCR step in the classic library preparation protocol [98].

1.1.2.2 Whole genome shotgun sequencing

With the rise of high-throughput second generation sequencing technologies [60,78,82] sequencing costs dropped extremely fast [51,99]. Multiplexing, i.e. combined sequencing of multiple samples in one sequencing run, was important for the cost reduction [100]. Tags are derived from short oligonucleotides with a distinct sequence. Specific oligonucleotides are added to DNA fragments of each sample during the sequencing library construction. These short oligonucleotides allow the binning of reads after the sequencing process thus reads can be assigned to a sample. As a result of low costs and high-throughput sequencing, whole genome shotgun (WGS) sequencing became the dominating strategy in genome sequencing projects. It replaced the previously applied hierarchical sequencing approach, which involved the cloning of genomic fragments into vectors like bacterial artificial chromosomes (BACs). In contrast, WGS relies on random fragmentation of multiple genome copies and following sequencing of these fragments in parallel. However, the quality of assemblies based on WGS reads was substantially inferior to the first reference genome sequences which were generated based on isolated Sanger sequencing of cloned genome fragments [101,102]. Depending on genome size, genome complexity, and available sequencing data WGS assemblies resulted in thousands [53], tens of thousands [21,103], or even hundreds of thousands of sequences [104]. To address different WGS assembly issues, there are approaches revisiting hierarchical BAC-based sequencing in combination with modern Illumina sequencing technologies to assemble highly repetitive genomes e.g. *Tritium aestivum, Saccharum* spp., and others [105–108]. Although the number of contigs is reduced in these approaches, the number of assembled sequences per genome remains high.

A high number of short contigs in an assembly comes with a high risk of including sequences derived from DNA contamination [102]. The presence of bacteria and fungi on plant leaves makes it almost impossible to extract clean plant DNA. As a result, numerous approaches were developed or adapted to address this issue bioinformatically. Examples are acdc [109], ProDeGe [110], Kraken [111], and various customized approaches based on sequence alignments [21,53,112,113]. Other major challenges for short read assemblers were repeats if the repeat length exceeds the length of reads or even the length of sequenced fragments [101,114]. For the same reason, gene duplications are likely to collapse in WGS assemblies [102]. This issue was avoided in first genome sequencing projects by cloning genome fragments and then resolving the sequence of only one fragment at a time.

1.1.2.3 Long reads of the third generation

The most important long read sequencing technologies are single molecule real-time (SMRT) sequencing provided by PacBio [60,76] and nanopore sequencing provided by ONT [115]. These long read technologies started the third generation of sequencing technologies and are currently dominating it.

SMRT sequencing relies on monitoring a polymerase fixed to the bottom of a zero-mode waveguide detector in real-time while this polymerase is integrating dye-labelled nucleotides into the newly synthesized DNA strand (Fig.3) [60,76,116,117]. Due to the small volume of each well the residence time of a fluorescent nucleotide is only sufficient for detection of an emitted light pulse if this nucleotide is incorporated [60]. Stochastically distributed insertions and deletions are the most frequent error type of this sequencing technology with an overall error rate of about 20% [60,118,119]. Reports of average read lengths in publications reached 20 kbp, while longest reads can even exceed 60 kbp [120,121]. PacBio claims that top read lengths of over 250 kbp can be achieved if the DNA quality is sufficient.

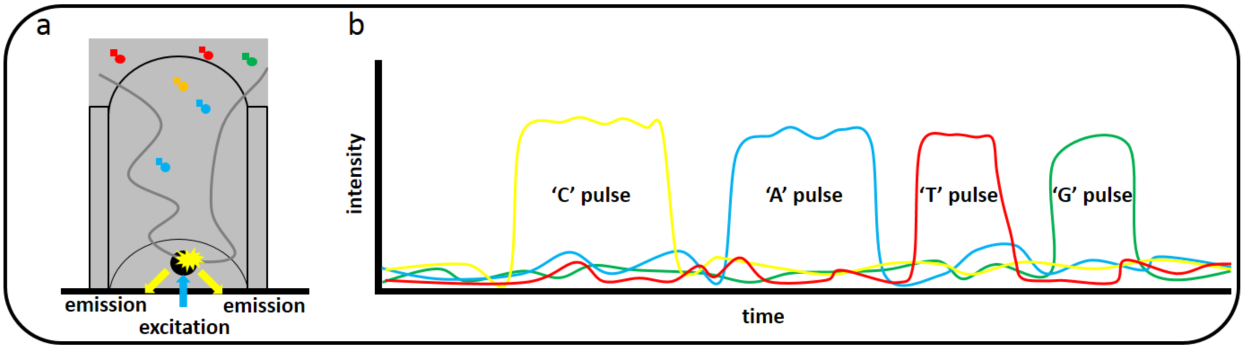

Fig.3: Single molecule real-time sequencing.

A DNA polymerase is fixed in a zero-mode waveguide (a). Only if a fluorescently labelled nucleotide is incorporated, the residence time in this well is sufficient to detect a signal. Even the incorporation of a single nucleotide results in detectable signals (b) which can be used to identify the respective base. Illustration concept is based on [76].

ONT provides an alternative technology for the generation of even longer reads [75,122]. Top lengths of sequenced DNA molecules are currently above 2 Mbp [123] thus the read length is mainly limited by the quality of the input molecules. Single molecules move through a pore in a membrane based on the electric charge of these molecules and cause changes in the ion flow through this pore by partly blocking it [124–127]. These changes in the ion flow are measured as current over the membrane. Current changes are specific to certain parts of the DNA [128,129], RNA [130,131], or even peptide [132] molecule being located in the pore at a certain time. It is currently assumed that six or even more nucleotides are affecting the signal at a given time resulting in a high number of k-mers which need to be distinguished [126,129,133]. This system is not restricted to determining the nucleotide sequence, but provides the opportunity to identify various modifications of nucleotides [134–138]. At the same time, these modifications pose an issue to the accurate sequence detection due to effects on the observed signal [133]. Controlled movement of a DNA strand at a constant speed through the nanopore is one of the biggest challenges [124,133]. Since homopolymers result in the same signal for several consecutive k-mers, measuring the duration of this signal is currently the best but still an erroneous approach to infer the homopolymer length [133,139–142]. Base callers like DeepNano [143], BasecRAWller [136], and Albacore (ONT) include neural networks to take information from adjacent k-mers into account [133]. It is possible to sequence just one strand (1D) or to sequence the forward and the reverse strand (2D or $1D^2$) [133,142,144–146]. Combining the sequencing results of both strands results in slightly more accurate reads [142,146]. Although the second generation of sequencing technologies enabled the generation of small genome sequences by single labs [51,53], especially the availability and portability of nanopore sequencing is currently revolutionising and democratising the field of genome assembly even further [122,147–151].

1.1.3 RNA-Seq

RNA-Seq, the massively parallel sequencing of cDNAs, is technically very similar to genomic sequencing workflows once the RNA of interest is reverse transcribed into cDNA. This technology revolutionized the field of gene expression analysis [77,152,153] and almost completely replaced array technologies [154,155]. Alternative splice variants of transcripts can be detected and transcript abundance can be quantified without prior knowledge about the sequence [2,77,156]. On the one hand, there is no longer an upper limit to the dynamic range of gene expression analyses, as the expression signal is inferred from counted reads [77,153]. On the other hand, lowly expressed transcripts can be detected as theoretically a single molecule would be sufficient to generate a countable read [157]. This comes with the additional benefit of a reduced amount of required sample material [77,157]. Quantification of transcriptional activity per gene is often performed by aligning reads to a genome or transcriptome sequence assembly and counting the number of reads assigned to each annotated gene or transcript, respectively. The alignment of RNA-Seq reads to an eukaryotic genome sequence requires dedicated split-read mappers like STAR [158] or HiSat2 [159] to account for the intron-exon structures of most genes. Since a high quality genome sequence is not always available, RNA-Seq is frequently applied to generate data for *de novo* transcriptome assemblies [160]. Transcriptome assemblies were used to discover candidate genes underlying a certain trait [113,161,162], to support gene prediction on a genome assembly [163,164], or to generate a reference sequence for transcript quantification [113,165].

Although RNA-Seq can be deployed to analyse all kinds of RNAs [157], many studies focus on mRNAs as these sequences encode proteins. Extreme differences in the abundance of different RNA types require isolation of the type of interest prior to sequencing. Enrichment of eukaryotic mRNAs is achieved via immobilized oligo-dT [166], oligo-dT priming during cDNA synthesis [167], selective hexamer priming during cDNA synthesis [168,169], or through depletion of rRNAs [170,171]. Oligonucleotides attached to magnetic beads can hybridize to rRNAs and enable controlled pulldown of this RNA type [170]. Since average mRNA lengths of many plant species are substantially exceeding 1000 nucleotides [7], resulting cDNAs are usually too long for direct sequencing. Therefore, an enzymatic or physical fragmentation step is needed [77]. While the fragmentation of cDNAs results in an enrichment of 3'-end fragments [77,152], the fragmentation of RNA would cause a depletion of 3'-end fragments [77,153]. Other applications focus on the analysis of small or non-coding RNAs [77,172]. Sophisticated protocols were developed to enable the strand-specific investigation of RNA to enrich the sequence with additional information about directionality of a molecule [173,174]. This information is important when quantifying the transcriptional activity of a gene. Observing sequences of mRNAs would indicate transcriptional activity, while sequences from complementary non-coding RNAs could indicate a repression of the gene of interest.

1.2 Genome assembly

Only fragments of a complete genome are represented in one read. Therefore, sophisticated software is required to reconstruct the genome sequence based on overlapping short reads [81,175,176].

1.2.1 Assembly of reads into contigs and scaffolds

Assembly results are usually continuous sequences (contigs) and scaffolds, which are composed of contigs separated by gaps of unknown sequence but estimated size. To avoid the inclusion of any artificial sequences like cloning vectors or adapters and to remove low quality reads, trimming of sequencing reads e.g. by trimmomatic [50] prior to the assembly is needed.

Assemblers evolved with the development of sequencing technologies. Assemblers for Sanger reads e.g. TIGR Assembler [177], Celera Assembler [178,179], CAP3 [180], and ARACHNE [181] expected long sequencing reads with a low error rate and a low sequencing depth [83,182].

Assemblers for second generation reads were mostly based on two general assembly paradigms: De Bruijn graph (DBG) [183–185] or overlap-layout consensus (OLC) [186] which have been nicely reviewed and explained before [160,187,188]. While the computation of overlaps between all reads in the OLC approach is a huge computational burden, it resolves many repeats [189]. However, the superior computational efficiency of DBG was the main reason for the application of DBGs in projects with large amounts of high quality short reads as generated by Illumina sequencers.

Frequently used DBG assemblers are Velvet [190], ALLPATHS-LG [191], SOAPdenovo2 [192], CLC [193], and SPAdes [194]. Platanus [114] is another example and was specifically developed for the assembly of highly heterozygous plant genome sequences. Newbler [78] is probably the most famous OLC assembler and was initially developed to assemble Roche/454 reads. While CABOG [83] is another OLC assembler, MaSuRCA [189] is combining OLC and DBG. However, there are also assemblers relying on different concepts e.g. the string graph assembler SGA [195].

Since it is often impossible to predict the best assembler for a given data set [196], it became best practice to empirically identify the best assembler and the best parameters by optimizing general assembly statistics [182]. The most important parameter for many assemblers is the k-mer size which depends on various factors e.g. the sequencing quality, the amount of reads, the read length, and the repeat content of the genome [197,198]. Some assemblers like Velvet and Platanus already come with support for the empirical identification of optimal assembly parameters [114,190].

The contiguity of WGS assemblies based on reads from second generation sequencing technologies can be improved through scaffolding. Tools like SSPACE [199] and SGA [195] utilize the information about approximate distances of paired-reads or mate pair reads to connect contigs and to estimate the size of gaps within scaffolds. After generation of scaffolds, gaps in the sequence can be filled using dedicated tools like GapFiller [200] or Sealer [201]. Although these gap sequences were not assembled in the first place, there might be reads which are actually covering these regions.

Long reads of the third sequencing generation pose computational challenges as assembly algorithms need to be optimized or even developed to take the characteristics of these new data types into accounts [73,119,202–205]. The major challenge of high error rates in long sequencing reads can be addressed by generating a high coverage [202]. As the distribution of sequencing errors is almost perfectly random within SMRT sequencing reads, an efficient detection and correction is feasible if sufficient coverage is given [202,205]. Multiple reads covering the same position of a genome of interest can be harnessed to infer the correct sequence at any position based on the sequence in the majority of all reads at and around this position. Correcting errors in ONT reads is more difficult and might not be accomplished by increasing the coverage as a component of the error is systematic [145,206]. In general, ONT reads have more deletions than insertions [207]. Even after application of various error correction tools [208–210], the reads possess a higher error rate than reads generated by sequencing technologies of the second generation. Despite these challenges, long reads of the third generation revolutionized the genome assembly field by enabling chromosome-level assemblies [54,126,203,211–213]. In respect to read properties, these new technologies display some similarity to Sanger sequencing. Therefore, it is not surprising that some of the new assemblers are in fact inspired by or even represent modifications of first generation assemblers [204]. Canu is based on the Celera Assembler [178,179] thus using an improved OLC approach [204]. It was reported before to be very efficient in the telomere assembly [214]. FALCON and FALCON-Unzip were developed to assemble haplophases of heterozygous species correctly [215]. Flye resolves repeats by selecting an arbitrary path through an A-Bruijn graph and corrects the resulting error-prone contigs in following steps [216]. Miniasm assembles contigs based on uncorrected reads in a time-efficient way [217]. Since long read processing is an extremely fast expanding field, there are many more tools under development. As both long read technologies display high error rates of up to 15% [208,209], polishing of generated assemblies can improve the sequence substantially [54,133,213]. Assembly polishing tools like Nanopolish for ONT read assemblies [203] and Arrow for SMRT sequencing read assemblies [202] utilize the random distribution of sequencing errors to correct an assembly by inspecting all mapped raw reads around a given position. Pilon [218] is not restricted to one long read technology and allows the polishing of assemblies e.g. based on mappings of Illumina reads [54,142,213]. Since an accuracy of 99.9% with insertions/deletions (InDels) being the main error type is not sufficient for gene prediction, polishing of raw assemblies with uniquely mapped Illumina reads is crucial [54,127,144,219,220].

1.2.2 Linkage information for high-level assembly scaffolding

After generation of contigs or scaffolds, anchoring of these sequences to chromosomes can be achieved through the incorporation of genetic markers [221,222] or by sequencing of fosmid, cosmid, yeast artificial chromosome, or BAC ends which provide long range linkage information [223]. Mapping of the read pairs from Hi-C data sets enables high level scaffolding [94,127,205]. Alternatively, BioNano Genomics and OptGen provide optical mapping information which can be incorporated into the scaffolding process [127,224,225]. Patterns of enzymatic restriction sites are investigated by electrophoretic analyses of fluorescently labelled DNA molecules which are up to several hundred kbp long [226,227]. The resulting patterns can be assembled into genome-wide maps which provide scaffolding information [121,228,229]. Many recent high quality assemblies of complex genomes rely on combinations of long sequencing reads and genetic linkage information derived from optical mappings [127,229,230].

A genetic map based on molecular markers can be used to achieve scaffolding on a very high level. The recombination between genetic markers is measured in centiMorgans (cM), the percentage of observed recombinations. There is a variety of marker types e.g. based on restriction fragment length polymorphisms [231], amplified fragment length polymorphisms [232–235], and simple sequence repeats [236]. Although genetic and physical maps are collinear, there are recombination hot spots and cold spots [237,238] which prevent direct correlation of genetic and physical distances. Nevertheless, genetic linkage supports the placement of assembled sequences resulting in high contiguity [212,213,219,239]. Genotyping-by-sequencing was recently applied for anchoring of assembled sequences [112].

1.2.3 Genome assembly validation

A huge variety of different sequencing technologies [60,72,81] and different assemblers requires careful assessment of the resulting assemblies to identify the best one [182,196,240–242]. Several competitions e.g. assemblathons were hold to characterize the performance of numerous assemblers on different data sets [182,196,242]. The results provide good hints towards suitable assemblers for a given sequencing data type. However, this assessment was limited to second generation sequencing technologies [182,242]. The increased pace of sequencing technology development and the corresponding development of novel assembly software makes it difficult to perform a benchmarking study which would be valid for a substantial amount of time.

In general, assembly quality assessment can harness the power of many orthogonal methods e.g. comparison of assembly statistics like N50 [242–244], inspection of read coverage depth after mapping reads against the assembly [245,246], assessment of mate distances in a mapping [246,247], and comparison against an existing reference sequence [53,191,241] or previously sequenced fragments of the same species [53,248]. There are trade-offs between certain properties e.g. high contiguity and correctness of an assembly [223,249–251]. While high contiguity, frequently measured as N50 [244], is generally desired to gain insights into the order and position of genetic features in a genome, the correctness of an assembly needs to be ensured. Mapping sequencing reads back to the final assembly is a very powerful approach to identify critical regions based on suspicious coverage values and positions of paired-end reads [247,252]. Collapses of multi copy genes or other repeats are indicated by substantially increased coverage values and broken pairs [247,252]. Miss-joints of contigs would lead to very low coverage values and a lack of spanning read pairs [247,252]. The completeness of assemblies can be assessed by looking for expected sequences like expressed sequence tags (ESTs) [53,253] or through comparison against a reference sequence [254]. In addition, genome size predictions based on sequencing reads [21] or biochemical assays like flow cytometry [255] can be compared against the assembly size. Tools like JellyFish2 [256], GenomeScope [257], and findGSE [258] estimate the genome size based on k-mer distributions in the sequencing reads. Since high error rates would bias such a prediction, the application of these tools is restricted to highly accurate reads of second generation sequencing technologies.

Since assembly assessment and validation is challenging and labour-intensive, dedicated tools were developed to support this task. QUAST [259] eases the comparison of different assemblies by calculation of statistics and optional comparison against a reference sequence. REAPR [247] inspects the read coverage depth and the distances of mates in a mapping of paired-end or mate pair reads. NucDiff [260] allows efficient comparison against an existing reference by utilizing NUCmer [261] to align genome sequences. Benchmarking Universal Single-Copy Orthologs (BUSCO) [262] can check an assembly for the presence of highly conserved genes which should be present in all genomes within a certain taxonomic group. Specific reference sequence sets for numerous taxonomic groups were generated to allow an optimal assessment [263]. Other approaches assess assembly quality based on the frequency of InDels in aligned regions where these differences are expected with a specific frequency [240].

1.3 Genome sequence annotation

After the generation of a high quality genome sequence assembly a structural annotation is needed to facilitate usage of this genomic resource. The prediction of protein encoding genes, RNA genes, transposable elements (TEs), and other repeats is a major challenge [187,264–267]. In general, gene prediction approaches can harness three types of information: i) sequence properties [268], ii) transcriptomic information e.g. RNA-Seq or ESTs [269,270], and iii) homology to other species [48,49,271]. The first information type is used in *ab initio* approaches, while the two latter approaches are hint-based [271].

1.3.1 Prediction of gene structures

Complex intron-exon structures in plant genomes prevent a simple search for protein encoding sequences [265,272]. Instead (generalized) hidden Markov models are frequently applied to account for species-specific gene model properties like intron size and codon usage [268,272–274] hence gene prediction can be performed *ab initio* without the incorporation of any hints [272]. Another approach is the transfer of information from closely related species through identification of sequence similarity thus an annotation is based on homology [48,49]. Frequently applied gene prediction tools are AUGUSTUS [272,275], various GeneMark derivatives [276–279], MAKER and MAKER2 [265,280], SNAP [281], Gramene [282], Gnomon [283], BRAKER [269], and GeMoMa [48,49]. Substantial improvements of the gene prediction are possible if RNA-Seq reads, ESTs, or sequences from closely related species [62,265,280,284,285] can be harnessed for the generation of hints. These sequences are mapped to the genome assembly to identify the positions of exons, introns, and especially the borders between exons and introns. Alignments of RNA-Seq reads against an assembly require the application of split-read aligners like STAR [158,286]. General alignment tools like BLAST [287] and BLAT [288] or dedicated tools like exonerate [289] can be applied to generate hints based on ESTs or sequences from related species. Annotations of the same sequence by multiple tools can be compared and even merged [290–292]. Hints from RNA-Seq reads can be used for the gene prediction process and additionally allow the selection of a final gene set based on transcription evidence [164]. While the *ab initio* prediction of protein encoding sequences is challenging, the prediction of features like UTRs or promoter sequences is even more difficult without hints [62,284,293].

1.3.2 Annotation of transposable elements

Annotation and classification of TEs is often omitted or poorly performed when annotating a genome sequence [294–296]. Since repeats and TEs account for substantial proportions of many genome sequences [23,297–299] and sometimes even have functional roles [299–307], both genomic feature types should not be ignored during the annotation process [295,308]. Numerous tools like RepeatScout [309] and RepeatMasker [310] are dedicated to the identification and annotation of repeat sequences and TEs [296,308]. Several tools were even combined into pipelines to harness individual strengths and compensate weaknesses [311], because no single tool was sufficient on its own [312,313]. Although there is little gain in masking repeats prior to the prediction of protein encoding genes [269], RepeatMasker is frequently deployed for this task [308]. Since the identification of TEs is challenging [296], well annotated TE sets of closely related species could be used to transfer the annotation and to flag predicted protein encoding genes as TE genes [213]. Due to the importance of TE annotation and the number of available tools and approaches, there is a huge need for a comparative benchmarking study to assess the performance of all tools on the same data set [296,308].

1.4 Re-sequencing projects and the diversity within species

High-throughput sequencing technologies enabled re-sequencing projects to investigate the genetic and genomic diversity within plant species (Fig.4) [40]. Although differences between accessions of the same species might be small, these differences can still cause variations in the outcomes of experiments [314]. A high number of accessions is available for *Arabidopsis thaliana* [53,254,315–327]. Some accessions were genotyped with focus on single nucleotide polymorphisms (SNPs) [322,328,329] or already subjected to Illumina sequencing [319–321,323,325,330]. In addition, varieties of various crop species were studied in similar re-sequencing projects [45,331–335]. Reads are mapped to a reference sequence using dedicated tools like Burrows-Wheeler Aligner (BWA)-MEM [336] or bowtie2 [337]. Large panels allow joint genotyping as provided by GATK [338,339]. Low confidence variants in multiple samples support each other and thus enhance the sensitivity of the variant calling process. As a result, sequence variants are identified with high reliability [330]. This investigation of 1,135 *A. thaliana* accessions revealed an average pair-wise difference of 439,145 SNPs [330] which results in one SNP in 271 bp. Calling variants based on long reads would efficiently identify substantially higher numbers of SVs than previously detected based on short reads [340–342]. Dedicated long read aligners like marginAlign [207], GraphMap [343], and PoreSeq [344] were developed to facilitate such variant detection approaches. However, recent improvements of the quality of sequencing technologies and advanced assembly algorithms might render reference sequences obsolete in the near future [51].

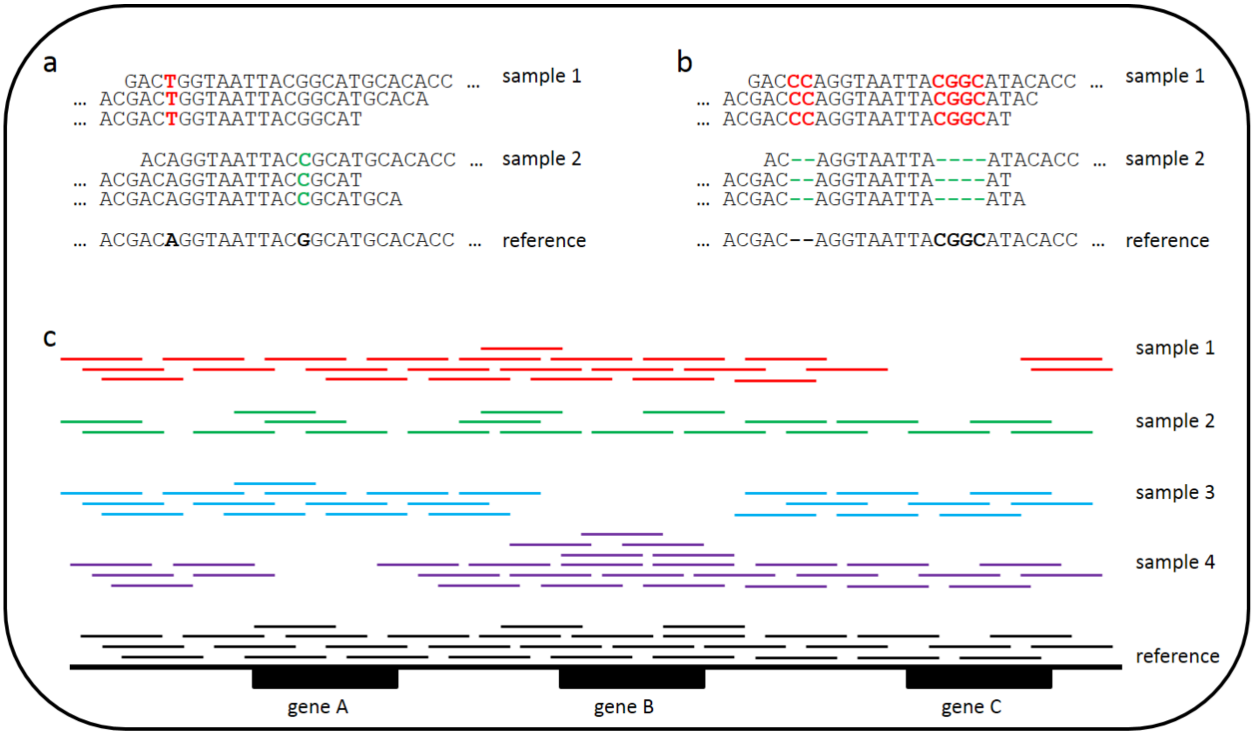
a
GACTGGTAATTACGGCATGCACACC ...
... ACGACTGGTAATTACGGCATGCACA
... ACGACTGGTAATTACGGCAT
sample 1
ACAGGTAATTACCGCATGCACACC ...
... ACGACAGGTAATTACCGCAT
... ACGACAGGTAATTACCGCATGCA
sample 2
... ACGACAGGTAATTACGGCATGCACACC ...
reference
b
GACCCAGGTAATTACGGCATACACC ...
... ACGACCCAGGTAATTACGGCATAC
... ACGACCCAGGTAATTACGGCAT
sample 1
AC--AGGTAATTA----ATACACC ...
... ACGAC--AGGTAATTA----AT
... ACGAC--AGGTAATTA----ATA
sample 2
... ACGAC--AGGTAATTACGGCATACACC ...
reference
c
sample 1
sample 2
sample 3
sample 4
reference
gene A
gene B
gene C

Fig.4: Detection of sequence variants.

Alignments of sequencing reads against a reference sequence reveal single nucleotide polymorphisms (a), insertions/deletions (b), and regions with coverage values deviating from the average (c). Cases of coverage deviation can be distinguished into presence/absence variations (PAVs) and copy number variations (CNVs). These variants are classified as PAVs if the sequence is unique, while repetitive sequences are considered CNVs.

1.5 *Arabidopsis thaliana* – a model organism for plant genomics

Arabidopsis thaliana (L.) Heynh. is THE model organism in plant genomics [1], general plant research [345–350], and plant systems biology [351]. Research on this plant was started in 1905 by Friedrich Laibach who collected first seeds around Limburg and from many other places in Germany [345,350,352,353]. The small genome size with a relatively low repeat content was beneficial for the generation of a high quality genome sequence through expensive and time-consuming BAC-based sequencing [1]. Assembled from Sanger sequencing reads [1], the Col-0 reference sequence remained the best plant genome sequence for almost two decades [120] and is still the best annotated one [2,349]. Many beneficial properties like a small size, short generation time, high number of seeds, and accessibility to genetic manipulation [348,354] facilitated the use of this plant species for functional genomics [2,349].

Some of the properties of *A. thaliana*, which made it a model organism in the first place, restrict its broader use due to substantial biological differences to many other plant species. Obviously, no model organism can be closely related to all species of scientific or economic interest (Fig.5). The transfer of knowledge from *A. thaliana* is generally more efficient over short phylogenetic distances. In contrast to most closely related species, *A. thaliana* is selfing instead of outcrossing and the chromosome number is reduced from eight to five [355]. The mating system of plants is of scientific interested for a very long time [356]. Selfing evolved several times independently in multiple plant species thus leading to a discussion about being an evolutionary dead end [357–360]. As an annual and herbaceous plant, it is not well suited for perennial plants and especially trees. In contrast to many other plant species, *A. thaliana* was assumed to be a non-mycorrhizal plant [361]. Although recent reports indicate that symbiotic interactions between *A. thaliana* and fungi do exist [362], important interactions between plants and fungi were studied in other models before. Despite all these limitations, *A. thaliana* is of high relevance for basic research. The extensive knowledge about the *A. thaliana* genome is also the basis for functional annotations of other plants [113] including crop species [363].

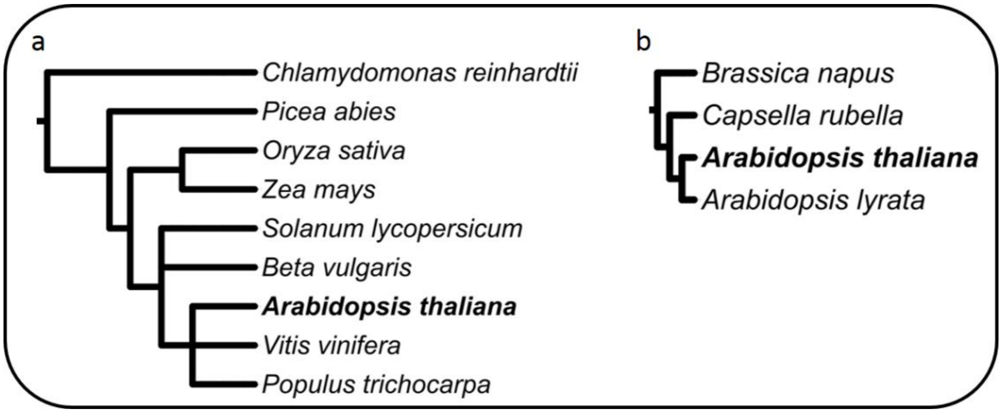

Fig.5: Phylogenetic position of *A. thaliana*.

The relative position of *A. thaliana* in a phylogenetic tree with important plant species (a) and with closely related species of the Brassicaceae (b) is displayed. Trees were constructed via phyloT [364] and iTOL [365].

1.5.1 Columbia-0, Landsberg *erecta*, and Niederzenz-1

Col-0 and Landsberg *erecta* (L*er*) are two accessions which are frequently used in research on *A. thaliana*. George Rédei generated L*er* by mutagenesis of Laibach's Landsberg strain and defined the Columbia (Col) accessions through single seed decent from the original Landsberg strain after observing that Landsberg was probably a mixture of different lines [366–368]. As a result, the genetically German accession Col-0 carries a name which points to the origin from the University of Missouri in Columbia where Rédei was working at that time [350]. L*er* was used in most studies due to a beneficial growth phenotype caused by the *erecta* mutation [350]. Nevertheless, Col-0 was selected for genome sequencing, because L*er* was expected to be substantially modified through the mutagenesis [1,350]. However, the importance of L*er* resulted in the publication of the chromosome-level genome sequence in 2016 [212]. The documented close relation between the two accessions with an available genome sequence suggests very similar genome sequences. Thus, additional *de novo* assemblies are needed to elucidate the intraspecific genomic diversity. Despite this very close relation between Col-0 and L*er*, a large inversion on chromosome 4 was identified between both accession and appears to occur in other *A. thaliana* accessions as well [212]. A possible explanation for this contradiction might be the heterogeneity of the initial Landsberg seed batch which was used by Rédei as origin of Col-0 and L*er*.

Fortunately, Nd-1 is independent of Col-0 and L*er* except for the geographic origin from central Europe. The name Niederzenz is assumed to indicate the geographic location where Laibach collected the first seeds [369]. Unfortunately, there is no village or town named Niederzenz thus the precise origin of Nd-1 remains unknown [369]. Several publications reported research on Nd-1 before the genome sequence was released [328,369–375] and Nd-1 was also included in the 1001 genomes project [376]. Additional motivation for the selection of Nd-1 as accession for a *de novo* sequencing project was the existence of recombinant inbred lines generated by crossing Nd-1 and C24 [372]. These lines were used to study biomass formation in *A. thaliana* [372] and provide a valuable resource for the investigation of other differences between both parents e.g. *BGLU6* which encodes a flavonol glucosyltransferase [375]. Although differences between Col-0 and Nd-1 exist and were described before, plants of both accessions cannot be distinguished optically under standard growth conditions in the greenhouse.

1.5.2 Genetics and genomics of *A. thaliana*

Genetic mapping approaches based on molecular markers were applied [377,378] e.g. to identify QTL [379] long before the first genome sequences were released. Famous are recombinant inbred lines which were developed to facilitate genetic studies in this model organism [380]. These lines were genotyped and allow an easy investigation of new phenotypic traits [378]. Investigations of genetic variations in *A. thaliana* have the potential to reveal new insights into development and physiology [381] and enhance the understanding of evolution [382]. Although genomic resources provide great potential for genome-wide association studies, the strong population structure of *A. thaliana* poses a challenge [383]. Intervals of variants in *A. thaliana* appear to be very small thus the resolution of genetic mapping can be in the single digit kbp range which is often equivalent to one or two genes [315,316,318,322].

The Col-0 reference sequence comprises approximately 120 Mbp [384] with 27,445 nuclear protein encoding genes included in the most recent Araport11 annotation [2]. Manually curated gene models and hints derived from numerous RNA-Seq data sets were incorporated in Araport11 [2]. Although the sequence is given as pseudochromosomes [384], there are a few completely missing [324,385–387] and collapsed [213,387,388] regions. Despite all efforts, the centromeric regions and nucleolus organizing regions (NORs) remained largely unassembled [1,54,212,213]. In total, there are still 29 large mis-assemblies [387] and over 90 gaps indicated by ‘N’ throughout the reference sequence [213,324]. Some of the most interesting genes were reported to be located in clusters of almost identical copies which are hard to assemble [378,389]. Although the Col-0 reference sequence is still of high quality compared to other assemblies, these issues are now addressed by long read sequencing technologies [54,120,211,213] which could improve the reference sequence through *de novo* assembly [390]. However, this reference sequence was crucial to investigate the evolution of species within the Brassicaceae [391,392], the biology of TEs [393–395], and genome evolution in general [391,392]. At least three whole genome duplications occurred during the phylogenic history of *A. thaliana* [391,396].

Paralogous gene copies, which originate from genome duplication, are called ohnologs [397]. There is still a substantial number of these ohnologs present in *A. thaliana*. While the genome duplication events probably took place about 7-12 million years ago (Mya), 47 Mya, and 124 Mya [391,398], the shift from outcrossing to selfing occurred only 150,000-1,000,000 years ago [123,399]. Associated with the shift from outcrossing to selfing could be the reduction in genome size as proposed before [400].

1.5.3 Molecular evolution

The rate of evolution i.e. the accumulation of variations over generations was studied in *A. thaliana* [320,323,401–403]. A strong bias towards conversion of G:C to A:T and an enrichment of mutations around the centromeres were reported as result of greenhouse experiments [402]. The average mutation rate was estimated to $7*10^{-9}$ substitutions per site per generation [402]. However, this substantial general excess of G:C to A:T conversion was not observed in natural strains [320]. This discrepancy can be explained by low frequency alleles of responsible SNPs in highly variable regions close to the centromeres [320]. A reduced selection pressure in *A. thaliana* compared to its closest sequenced relatives *A. lyrata* and *A. halleri* was reported to enhance the rate of protein evolution [404,405]. This reduced selection pressure could be caused by the shift from outcrossing to selfing, because the effective population size was reduced [405–407] and therefore an accelerated rate of protein evolution can be assumed [408]. A lower purifying selection and a higher mutation rate is assumed to increase the rate of pseudogenization [404,405].

1.5.4 Genome size of *A. thaliana*

Although the first genome sequence of *A. thaliana* was provided almost 20 years ago [1], the precise genome size is still unknown. While the common ancestor of all Brassicaceae had an estimated genome size of 500 Mbp [409] distributed over eight chromosomes, there are major differences between the genome structures of derived species [410,411]. In contrast to other *Arabidopsis* species, *A. thaliana* has only five chromosomes and an estimated genome size of 130-150 Mbp [1,53]. The genome size difference between *A. thaliana* and *A. lyrata* was partly attributed to small InDels, differences in heterochromatic regions, and differences in the number of TEs [412]. Comparison with *Capsella rubella* revealed the TE differences between the *Arabidopsis* species as a derived characteristic in *A. lyrata* [413]. There are even reports of intraspecific genome size differences in *A. thaliana* with Col-0 displaying a relatively small genome [258,324,414]. The number of rDNA repeats, which encode the 45S rRNA and are located in the NORs, were also identified as important sources for genome size differences [324,415]. Previously, genome sizes were investigated over multiple decades while the resolution of applied technologies increased. Deployed methods included reassociation kinetics [416], quantitative gel blot hybridization [417], Feulgen photometry [418], flow cytometry [255,419], and k-mer-based calculation to harness the power of second generation sequencing technologies [21,53,257]. However, recent developments in sequencing technologies promise complete genome assemblies as the ultimate method to assess the genome size precisely.

1.5.5 Transposable elements in *A. thaliana*

Besides polyploidization, TE amplification is one of the major forces contributing to the genome size [23,420,421]. In comparison to other plant genomes, the TE and repeat contribution to the known *A. thaliana* genome is relatively small with only 10-30% [1,395]. TEs in *A. thaliana* Col-0 were annotated in 2008 [393] and despite some issues [394,395,422] this annotation was never updated [2,384]. Since TEs are generally less active in selfing plants [395,423], observed losses of TEs in *A. thaliana* compared to outcrossing relatives like *A. lyrata* are expected. Re-sequencing projects revealed already that up to 80% of all annotated TEs appeared to be fragmented or deleted in at least one accession [320].

1.5.6 Gene set of *A. thaliana*

The minimal set of genes necessary for a plant to survive under controlled conditions or in the natural environment is still unknown [246,424]. While some genes might not be necessary for survival, these genes could still be beneficial or even necessary under specific conditions [246]. *A. thaliana* is not an ideal model to address these questions since many functionally redundant ohnologs are still present as a result of the ancestral genome duplications. Nevertheless, the comprehensive annotation of 27,445 protein encoding nuclear genes in the *A. thaliana* genome sequence [2] is very beneficial for gene set investigations. While copy number variations just alter the gene dose, presence/absence variations (PAVs) distinguish between wild type and knock-out.

Previous studies reported 620 *A. thaliana* genes which are involved in the seed development or physiology thus causing visible differences to the wild type when knocked out [425]. Sets of 130 and 60 essential genes were identified in the female gametophyte development and male gametophyte development, respectively [424,426]. While these results are derived from knock-out experiments, the natural diversity of *A. thaliana* provides the material to classify genes based on presence/absence in various accessions as ‘core’ or ‘non-core’. Only genes present in all accessions belong to the core gene set. Although this set of core genes is not necessarily identical with the set of essential genes, a strong overlap can be expected. Genes absent from at least one accession can be considered to be dispensable, because plants of one accession are apparently able to survive without these genes. Re-sequencing projects revealed copy number variations (CNVs) and PAVs between numerous accessions and the reference sequence, which involved several hundred genes [320,427]. In total, 26,373 genes were identified as core genes of 19 *A. thaliana* accessions and 11,416 additional ones were classified as accessory genes [428]. Not just the absence of genes is informative to narrow down the core gene set, but also reports about genes with sequence variants likely to render a gene functionless. In total, 4,263 genes with a premature stop codon in at least one accession were identified during a re-sequencing project, but the false positive rate of this process is high [320].

Besides the search for a minimal gene set, the identification of genes unique to one accession is an important contribution to the pan-genome of *A. thaliana*. The pan-genome comprises all genes or even non-genic sequences which are present in at least one member of a species [429,430]. Comprehensive knowledge of the pan-genome is necessary to understand the genetic and genomic diversity within a species [430]. Assembly quality, annotation quality, detection of orthologs, and the selection of appropriate samples are main factors determining the quality of pan-genome analyses [430]. Differences in the gene sets of individuals were previously proposed as the basis of heterosis effects [431] which are important in plant breeding.

1.5.7 Transcriptomics of *A. thaliana*

Gene expression in this model organism was assessed by RT-qPCR [432–435] and array technologies for years [436–439] resulting in comprehensive expression databases [440–444]. The rise of high-throughput sequencing technologies enabled the investigation via RNA-Seq [174,386] thus facilitating the detection of transcripts in a reference-independent way [77]. The most recent annotation of the Col-0 reference sequence is based on a set of diverse RNA-Seq data sets and focused on the annotation of numerous RNA genes [2]. *De novo* transcriptome assemblies based on RNA-Seq reads revealed sequences which could not be mapped to the reference genome sequence [445,446]. Only a small number of novel genes were detected [445,446], but these reports indicate that not all expressed genes are represented in the current Col-0 reference sequence.

1.6 Splicing and splice sites

Plant genes harbour an average of 4.5 introns per protein encoding gene [7], which separate the exons and require a removal from transcripts prior to translation [447–450]. Splicing, i.e. the removal of introns from primary transcripts, involves five snoRNAs and over 150 proteins which are associated in the spliceosome [451]. Different types of introns are recognized and removed by the U2 [452] or the U12 [453] spliceosome, respectively. Discussions about the classification of introns, potential additional spliceosomes [454], and minimal intron sizes [455–457] are still ongoing. Specific binding of the spliceosome and proper removal of introns require highly conserved sequences around the splice sites [458–460]. The terminal dinucleotides of introns are highly conserved: GT at the 5'-end and AG at the 3'-end on the DNA level [7,62,461]. These GT-AG splice site combinations are named canonical. There are also rare cases where terminal dinucleotides deviate from the canonical GT-AG sequence resulting in so called non-canonical splice site combinations [3,7,62]. The major non-canonical splice site combinations GC-AG and AT-AC account on average for 1.3% of all splice sites in plant genomes [7]. Minor non-canonical splice site combinations display all other nucleotide combinations at a much lower average frequency of approximately 0.1% (Fig.6) [7].

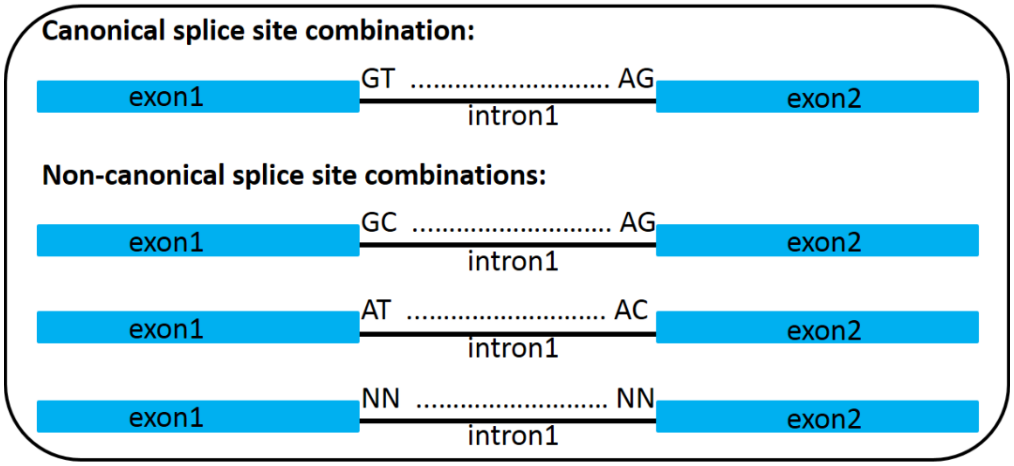
Canonical splice site combination:
exon1
GT .... AG
intron1
exon2
Non-canonical splice site combinations:
exon1
GC .... AG
intron1
exon2
exon1
AT .... AC
intron1
exon2
exon1
NN .... NN
intron1
exon2

Fig.6: Splice site combinations.

Besides the canonical GT-AG splice site combination, there are two major non-canonical splice site combinations: GC-AG and AT-AC. In addition, all other dinucleotide combinations might occur as minor non-canonical splice site combinations (NN-NN), but the frequency drops with divergence from the canonical sequence. Although the actual splicing process modifies RNA, all sequences in this thesis refer to the corresponding DNA sequence.

1.7 Objectives

Arabidopsis thaliana is well established as a model organism for many years. However, only a single genome sequence at chromosome-level quality was described in the literature at the beginning of this work.

Therefore, the first objective was to generate a *de novo* genome assembly and a corresponding annotation of the *A. thaliana* accession Niederzenz-1 (Nd-1). Numerous comparative genomic analyses are enabled through the availability of the here presented highly contiguous genome sequence. Synteny, structural variants, and copy number variations between *A. thaliana* accessions are investigated. Novel sequences are inferred from this *de novo* assembly thus contributing to the pan-genome of *A. thaliana*. An independent high quality assembly can also facilitate the correction of errors in the Col-0 reference sequence.

The second objective was to investigate non-canonical splice sites in *A. thaliana*. These splice sites evade *ab initio* gene prediction causing erroneous gene structures. As a model plant *A. thaliana* is the perfect system to establish methods for an improved gene prediction and for the investigation of non-canonical splice sites in other species.

The third objective was to transfer knowledge about non-canonical splice sites in *A. thaliana* and methods for the investigation of these splice sites to other plants. Since existing knowledge about this topic was sparse, a comprehensive investigation of non-canonical splice sites was necessary to shed light on this topic and to provide resources for future studies. The analysis of over 120 plant genome sequences and annotations requires automation. Implementing the analysis workflow in Python scripts provides scalability and transferability.

2 Results

The results of this work are described within the following chapters. Summaries of the four research items are presented with figures, tables, and the corresponding captions coming from these items.

First, the *A. thaliana* Niederzenz-1 (Nd-1) *de novo* genome sequence assembly based on second generation sequencing data is presented [53]. The Nd-1 genome sequence was compared against the Columbia-0 (Col-0) reference sequence to identify small sequence variants and PAVs.

Next, insights into non-canonical splice site combinations in Col-0 are presented. Based on this knowledge an improved annotation of the Nd-1 genome sequence assembly is presented. This improvement was achieved through consideration of non-canonical splice site combinations during the gene prediction process [62].

Afterwards, a *de novo* genome sequence assembly based on SMRT sequencing is presented together with an extended comparison against the Col-0 reference sequence [213]. Large structural variants (SVs) between Nd-1 and Col-0 were revealed by this assembly. Copy number variations (CNVs) and presence/absence variants (PAVs) across numerous *A. thaliana* accessions were identified by mapping of reads against the assembly and investigation the resulting coverage values.

Fourth, the analyses of non-canonical splice sites are extended to all annotated plant genome sequences and supplemented with support from transcriptomics [7]. Methods developed based on *A. thaliana* are adjusted and optimized for application on various plant genome sequences to enhance automation of the analyses.

2.1 *De novo* genome sequence assembly of *A. thaliana* Nd-1

Re-sequencing of various *A. thaliana* accessions was performed for years [120,212,254,324], but this approach is mostly limited to the detection of small sequence variants [254,320,324]. Reference-independent *de novo* assemblies are needed to resolve larger insertions, to detect SVs, to identify PAVs, and to assess synteny [102,254]. Therefore, the Nd-1 genome was analysed using various second generation sequencing technologies.

2.1.1 The *A. thaliana* Nd-1 assembly

Approximately 120 fold coverage of Illumina paired-end reads and additional linkage information from mate pair reads was used for the Nd-1 genome sequence assembly. The resulting assembly comprised 5,197 scaffolds with an N50 of 0.59 Mbp (Table 1). The Nd-1 genome size was estimated to 146 Mbp based on k-mer distributions of reads. This Nd-1 assembly covers approximately 99.8% of the Col-0 reference sequence, while a read mapping covered only 96%. SVs within the assembled regions are likely to explain this difference. Most Nd-1 scaffolds are mapped close to the peri-centromere sequences on the five Col-0 pseudochromosomes. TEs were identified as a challenge to the assembler, because 65% of the mapped contig ends are matching an annotated TE in the Col-0 reference sequence. A total of 28,670 protein encoding genes were predicted in this assembly. The encoded proteins were assigned to Col-0 proteins via reciprocal best BLAST hits (RBHs) to transfer the comprehensive functional annotation of the Col-0 annotation to Nd-1. An analysis of the positions of the identified 22,178 RBHs revealed strong synteny between Nd-1 and Col-0 (Fig.7). Genes in RBH pairs with non-syntenic positions were partly caused by close paralogs which prevent the detection of proper orthologs.

Table 1: Assembly statistics.

Metrics of the Nd-1 genome sequence assembly generated via CLC Genomics Workbench before and after application of SSPACE, GapFiller and subsequent RBH-based manual improvement.

parameter	CLC assembly	scaffolded	gaps filled	polished
number of scaffolds	10,057	5,201	5,201	5,197
total number of bases	113,939,710	117,144,260	117,816,107	116,846,015
average scaffold length	11,329 bp	22,523 bp	22,652 bp	22,483 bp
minimal scaffold length	500 bp	500 bp	500 bp	500 bp
maximal scaffold length	445,914 bp	3,176,818 bp	3,190,961 bp	2,967,516 bp
GC content	35.98%	35.98%	35.95%	35.95%
N25	102,863 bp	1,299,823 bp	1,304,062 bp	1,211,412 bp
N50	52,252 bp	709,626 bp	713,021 bp	589,639 bp
N75	22,586 bp	214,378 bp	215,617 bp	174,007 bp
N90	7,163 bp	42,960 bp	43,285 bp	40,994 bp

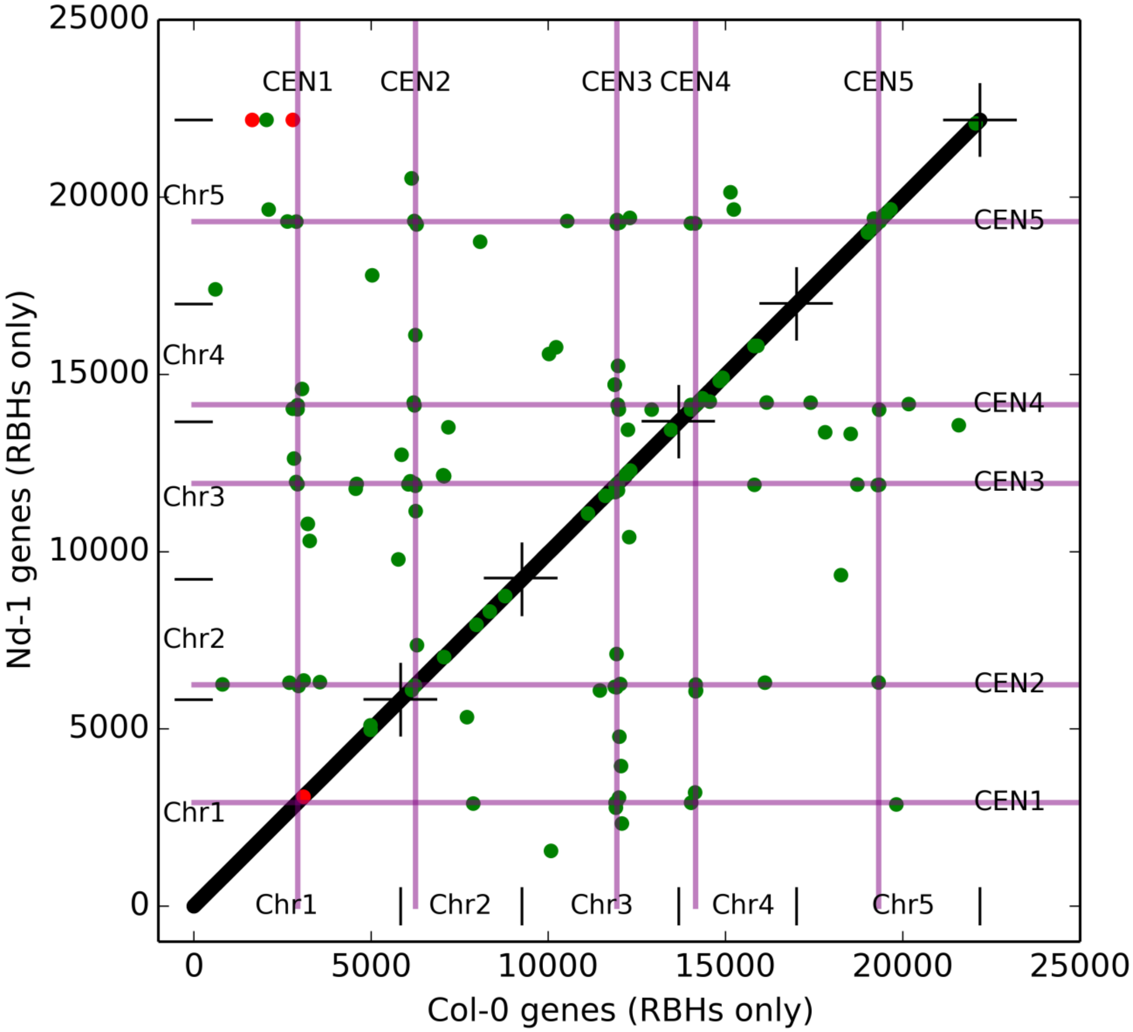

Fig.7: Synteny between Nd-1 and Col-0 based on reciprocal best BLAST hits.

All five pseudochromosomes of the two genome sequences were ordered by their number to provide the x (Col-0) and y (Nd-1) axes of the diagram. Positions of each RBH pair in the two genome assemblies were plotted, resulting in a bisecting line formed from black dots representing perfectly matching RBH pairs. RBH gene pair positions deviating from a fully syntenic position, i.e. the outliers, are represented by green dots for RBH pairs with ambiguous best hits in RBH pair identification, and by red dots for RBH pairs with deviating (non-syntenic) gene positions. Since two red dots overlap each other, only three locations are visible. Positions of the centromeres (CEN1 to CEN5) are indicated by purple lines. Ends of pseudochromosomes (telomeres) are indicated by short black lines at the bisectrix (forming crosses) and on both axes. Formally, the unmapped fraction of

Nd-1 contigs is appended after pseudochromosome 5, but this sequence of about 134 kbp in length becomes invisible due to the limited resolution of the figure.

2.1.2 Small sequence variants

The generated sequencing data were also subjected to a read mapping against the Col-0 reference sequence to enable variant detection. A total of 485,887 identified single nucleotide polymorphisms (SNPs) and insertions/deletions (InDels) were functionally annotated to find genes with relevant differences. While 314 genes displayed premature stop codons, 117 genes lost the stop codon, and 1,228 additional genes displayed frameshifts. The genome-wide distribution of the small sequence variants did not reveal specific clusters and did not show substantial differences between the chromosomes (Fig.8).

When comparing InDel lengths between protein encoding sequences and other regions, a substantial difference in the distributions emerged (Fig.9). Protein encoding regions displayed an increased frequency of InDel lengths that are a multiple of three as these do not shift the reading frame.

Previously reported effects in *RRS1* (At5g45260) [371,462], *MYB114* (At1g66380) [374], and *BGLU6* (At1g60270) [375] were recovered in this analyses thus supporting its validity and value.

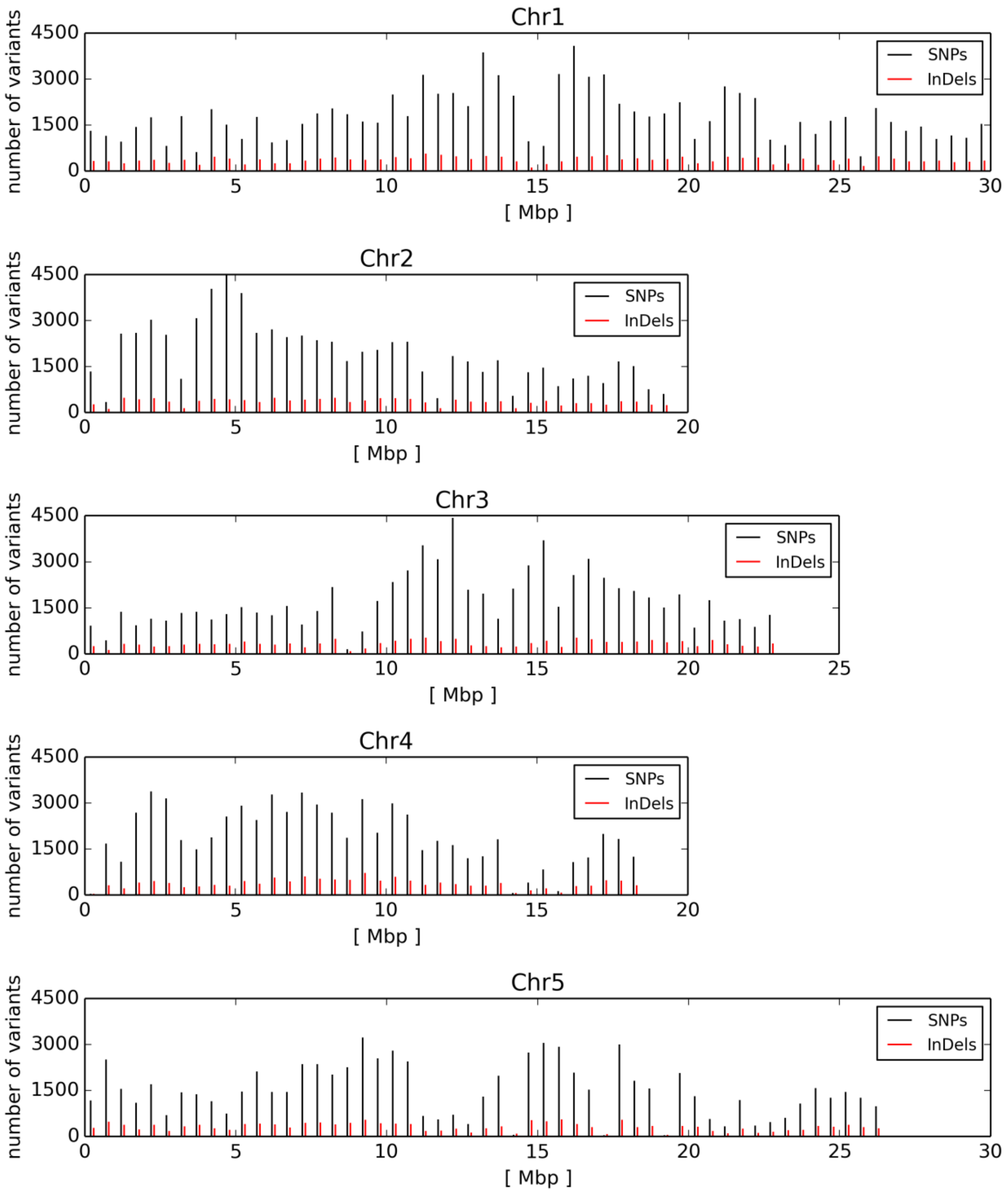

Fig.8: Genome-wide distribution of small sequence variants.

Numbers of SNPs (black) and InDels (red) in a given interval on the chromosomes are shown. Both variant types were identified using GATK and CLC Genomics Workbench as described in the method section [53]. The overlap of both tools was considered as the best choice.

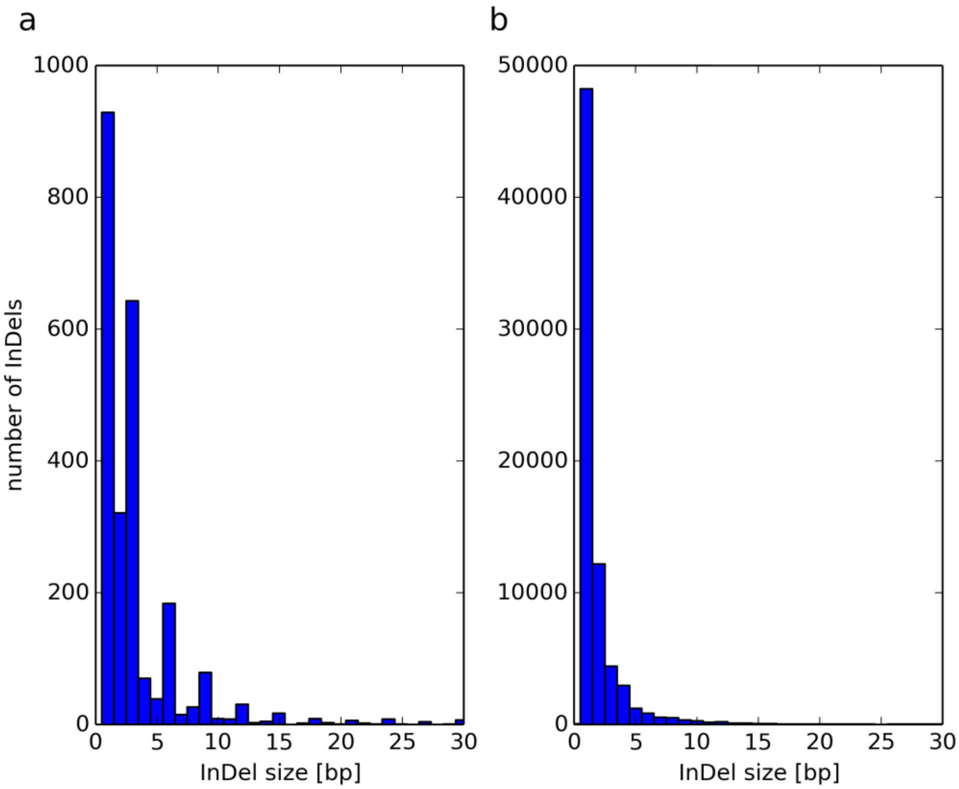

Fig.9: Insertion/deletion size distributions.

Most frequent InDel sizes differ between protein encoding sequences (a) and non-coding sequences (b). InDel lengths that are a multiple of three are much more common in protein encoding sequences.

2.1.3 Presence/absence variations

Thousands of PAVs and highly divergent regions were identified between Nd-1 and Col-0 through substantial differences in the read mapping coverage. The PAV size ranged up to 53 kbp with a combined size of 5.5 Mbp. After validation via sequence alignment, randomly selected candidates were validated via PCR (Fig.10).

The Nd-1 assembly revealed a previously described modification of the *FLOWERING LOCUS M* (At1g77080) [373]. In addition, two copies of *SEC10* (At5g12370) which are collapsed in the Col-0 reference sequence [388] were correctly resolved in this assembly.

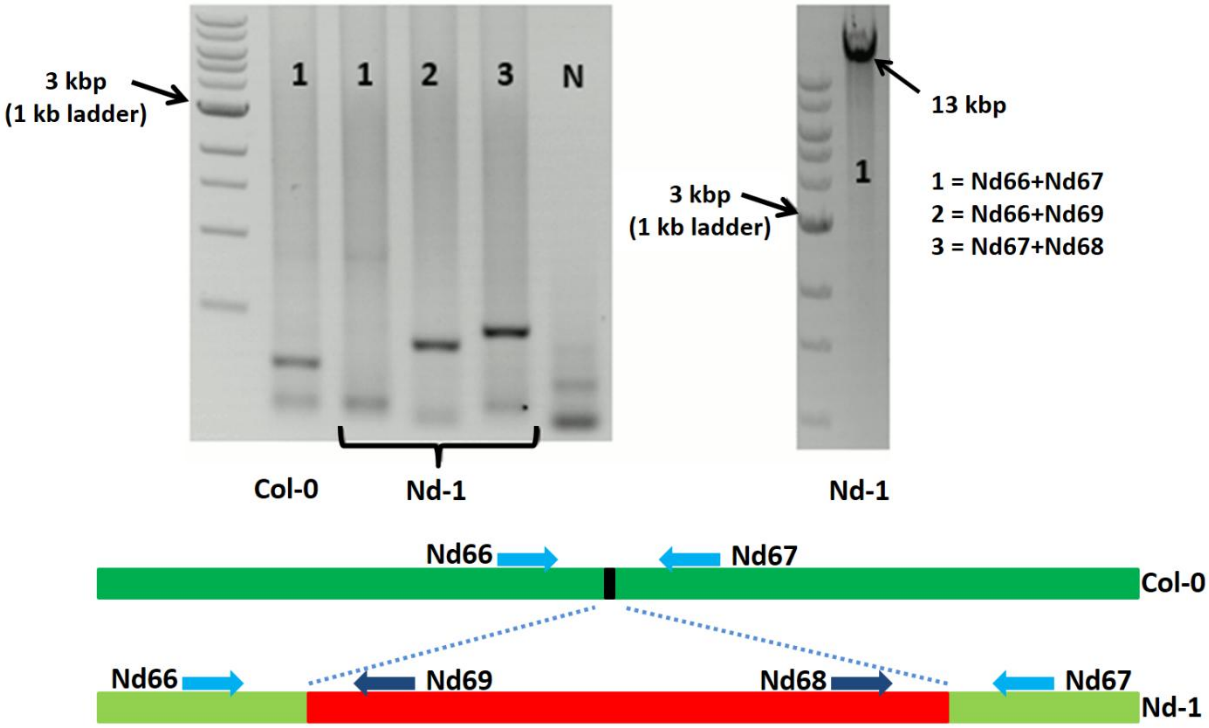

Fig.10: Validation of an insertion in Nd-1 via PCR.

The concept is visualized by using a PAV of about 13 kbp in length that is present in Nd-1 and absent from Col-0 as an example. This figure shows the primer positions used for experimental validation (bottom). Outer primers (Nd66 and Nd67) were used for standard PCR on genomic DNA of Col-0 and Nd-1 (gel picture of amplicons, top left) and for long range PCR on genomic DNA of Nd-1 (top right). Inner primers were used for amplicon generation in standard PCR with genomic DNA of Nd-1.

2.2 Consideration of non-canonical splice sites improves gene prediction

Terminal dinucleotides in intron sequences are highly conserved, because these sequences are crucial signals in the mRNA splicing process. Despite strong conservation of the canonical GT at the 5’ splice site and AG at the 3’ splice site, some variants of this splice site combination occur at low frequencies [4,463]. Besides the major non-canonical splice site combinations GC-AG and AT-AC, there are numerous combinations of minor non-canonical splice sites occurring at very low frequencies [4,463]. These exceptions pose a severe challenge to *ab initio* gene prediction approaches, which try to identify gene structures based on sequence properties [464,465]. The low frequency of non-canonical splice site combinations would cause a substantial number of false positive splice site predictions if considered [466,467]. In addition, the number of possible gene models would increase extremely if all dinucleotides would be considered as potential splice sites [466,467]. Therefore, *ab initio* gene predictions are only identifying canonical splice site combinations resulting in erroneous predictions of genes with *bona fide* non-canonical splice sites [466,467].

2.2.1 Non-canonical splice sites in *A. thaliana*

The Araport11 annotation [2] of Col-0 contains 1,267 genes with non-canonical splice sites in the representative transcript i.e. the transcript with the longest CDS. While 98.9% of all splice site combinations are the canonical GT-AG, there are 1.0% GC-AG, and 0.1% AT-AC major non-canonical splice site combinations. Consequently, the remaining 0.1% (88 introns) are minor non-canonical splice site combinations.

The actual usage of these non-canonical splice sites was validated for *FGT1* (At1g79350), *AGY1* (At4g01800), and *PPI1* (At4g27500) via amplicon generation from cDNA and following Sanger sequencing. Independent Illumina sequencing data were used to validate the corresponding DNA sequences.

2.2.2 Improved gene prediction

To harness the full power of the manually curated annotation of Araport11, representative transcript sequences were mapped to the Nd-1 assembly [53] to generate hints for the gene prediction. Hints for exons or exon-intron borders, respectively, are required to enable the prediction of non-canonical splice sites. Again, AGUSTUS was applied for the prediction of gene structures based on the generated hints and information about expected minor non-canonical splice site combinations. The resulting 30,834 protein encoding gene models matched 91.2% of the CDS features in the *ab initio* annotation and 50.2% of the UTRs. The distribution of 99.0% canonical splice site combinations, 0.8% GC-AG major non-canonical splice site combinations, 0.05% AT-AC major non-canonical splice site combinations, and

0.15% (206) minor non-canonical splice site combinations is similar to Col-0. The total of 1,256 genes with non-canonical splice site combinations in Nd-1 is almost perfectly matching the number of 1,267 genes with non-canonical splice site combinations observed in Col-0. The importance of hints in the gene prediction in respect to non-canonical splice sites can be seen when looking at the RBHs of *FGT1*, *AGY1*, and *PPI1*. For example, intron20 of *FGT1* in Nd-1 displays non-canonical splice sites at both terminal ends (Fig.11). Therefore, the *ab initio* gene prediction is skipping the exon20 completely (Fig.11). In general, the identification of substantially more RBHs between Araport11 and the GeneSet_Nd-1_v1.1 compared to the previous *ab initio* annotation GeneSet_Nd-1_v1.0 indicates an increased annotation quality. Since Araport11 contains the manually improved annotation of Col-0, it can be considered a gold standard.

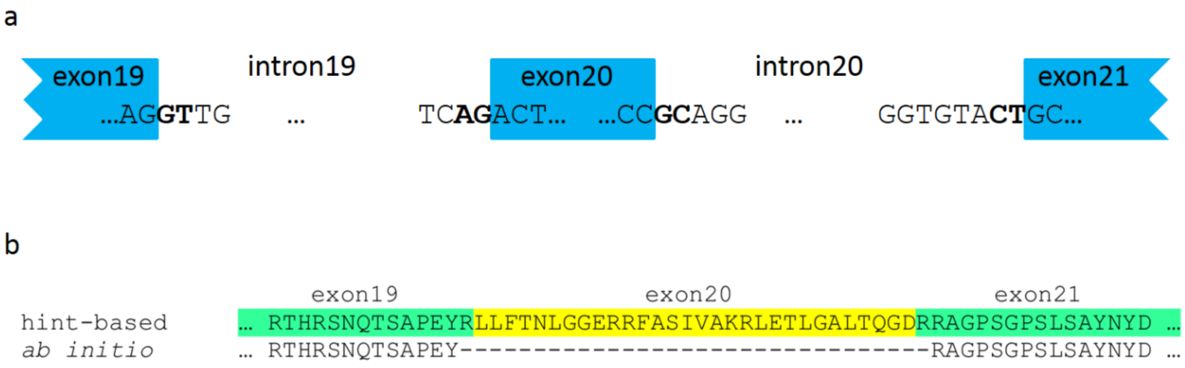
a
exon19
intron19
exon20
intron20
exon21
...AGGTTG
...
TCAGACT...
...CCGCAGG
...
GGTGTACTGC...
b
exon19
exon20
exon21
hint-based
ab initio
… RTHRSNQTSAPEYRLLFTNLGGERRFASIVAKRLETLGALTQGDRRAGPSGPSLSAYNYD …
… RTHRSNQTSAPEY--------------------------------RAGPSGPSLSAYNYD …

Fig.11: Representative gene structure of missed non-canonical splice sites in the *ab initio* gene prediction on Nd-1.

Gene structure of the At1g79350 RBH in the hint-based gene prediction (GeneSet_Nd-1_v1.1) on the Nd-1 genome sequence is displayed (a). The non-canonical splice sites were missed in the *ab initio* gene prediction leading to a skipping of exon20 (highlighted in yellow) (b).

2.3 Chromosome-level assembly reveals structural variants between Nd-1 and Col-0 and facilitates pan-genomic analyses

One important motivation for *de novo* assemblies of genome sequences for species with an available reference sequence is the detection of structural variants (SVs) [102]. Whole genome shotgun (WGS) assemblies based on short reads already revealed some variants up to several kbp between Col-0 and Nd-1 [53]. However, the limited assembly contiguity hampered the identification of large structural rearrangements. N50 values of *A. thaliana de novo* genome sequence assemblies were usually far below 1 Mbp [53,254]. With long sequencing reads generated by technologies of the third generation, assembly contiguity improved substantially [120,202,211] making chromosome-level assemblies possible at least for model organisms like *A. thaliana* [54,212].

2.3.1 Assembly based on SMRT sequencing reads

A Canu [204] assembly based on SMRT sequencing reads (Ath-Nd-1_v2c) was selected as representative Nd-1 genome sequence assembly after evaluating results of different assemblers (Table 2). Assuming a genome size of 150 Mbp the coverage of this data set was 112 fold. Although it was not the focus of this work, differences between the tested assemblers were observed. While FALCON was unable to resolve sequences close to some telomeres, these regions were included in the Canu assembly Ath-Nd-1_v2c. The total assembly size exceeds the original Col-0 reference sequence [1] by about 3 Mbp. Substantial improvement of the assembly contiguity over the previous Ath-Nd-1_v1 assembly [53] is indicated by the N50 of 13.4 Mbp. In addition, AthNd-1_v2c resolves 1,744 additional TEs compared to Ath-Nd-1_v1 [53]. Nd-1 contigs were placed and orientated based on genetic linkage information, where possible. This linkage information was derived from genotyping F2 plants of reciprocal crossings of Nd-1 and Col-0. Unanchored small contigs were placed based on the Col-0 reference sequence and all contigs mapped to the same chromosome were joined into a pseudochromosome. Ath-Nd-1_v2c bridges several regions where the Col-0 reference sequence is interrupted by gaps. Compared to Ath-Nd-1_v1, 6.9 Mbp additional sequence mostly close to the centromeres is included in Ath-Nd-1_v2c. Assembly completeness is also indicated by the presence of telomeric repeats at the end of most pseudochromosomes (Fig.12). While the WGS assembly Ath-Nd-1_v1 was not able to resolve nucleolus organizing repeats (NORs) automatically, there are several repeat units represented in Ath-Nd-1_v2c (Fig.12). BUSCO [262] detected all benchmarking genes which are also found in the Col-0 reference sequence thus providing additional evidence for a high quality assembly.

Table 2: Nd-1 *de novo* assembly statistics.

Metrics of assemblies of the Nd-1 nucleome sequence generated by Canu, FALCON, miniasm, and Flye, respectively. All described assemblies are the final version after polishing.

parameter	Ath-Nd-1_v2c	Ath-Nd-1_v2f	Ath-Nd-1_v2m	Ath-Nd-1_v2y
Assembler	Canu	FALCON	Miniasm	Flye
number of contigs	69	26	72	44
total number of bases	123,513,866	119,540,544	120,159,079	116,964,092
average contig length	1,790,056 bp	4,597,713 bp	1,668,876 bp	2,658,274 bp
minimal contig length	50,345 bp	86,055 bp	50,142 bp	53,207 bp
maximal contig length	15,898,009 bp	15,877,978 bp	14,338,505 bp	14,857,908 bp
GC content	36.14%	36.04%	36.07%	36.01%
N25	14,369,729 bp	14,534,675 bp	11,880,610 bp	12,510,540 bp
N50	13,422,481 bp	9,302,209 bp	8,595,164 bp	10,607,548 bp
N75	8,555,326 bp	6,666,836 bp	3,513,050 bp	6,001,858 bp
N90	2,928,047 bp	2,829,734 bp	1,430,525 bp	2,524876 bp

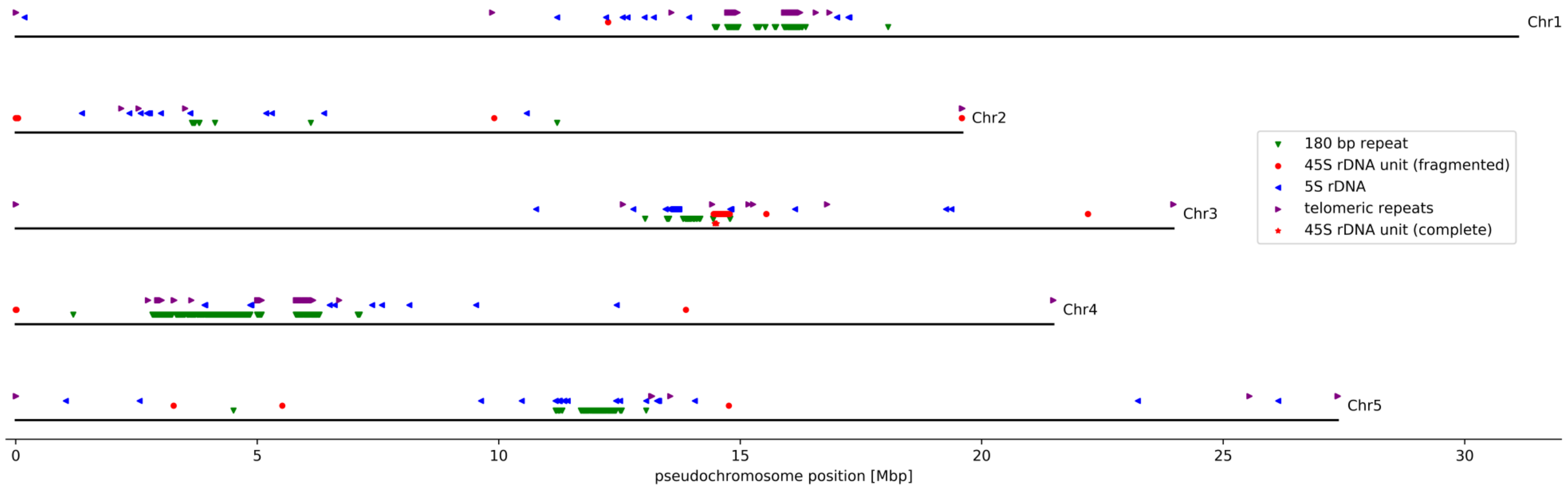

Fig.12: Nd-1 genome structure.

Schematic pseudochromosomes are shown in black with centromere repeat positions in green. Red dots indicate positions of 45S rDNA fragments and an orange star represents complete 45S rDNA transcription units. Blue triangles indicate the positions of 5S rDNAs. The positions of telomeric repeats are shown by purple triangles.

2.3.2 Large structural variants

The most striking difference between Col-0 and Nd-1 is a 1 Mbp inversion in the north of the chromosome 4 (Fig.13). This inversion allele in Nd-1 is slightly different from the 1.2 Mbp inversion allele reported previously in L*er* [212]. As reported for the L*er* allele [212], a repression of recombination events in this region was also observed for Nd-1 while studying genetic linkage. In addition, there is a translocation on chromosome 3 effecting At3g60975-At3g61035. Several smaller SVs are clustered around the centromeres. Since these regions are highly repetitive, assembly or alignment errors could influence this observation.

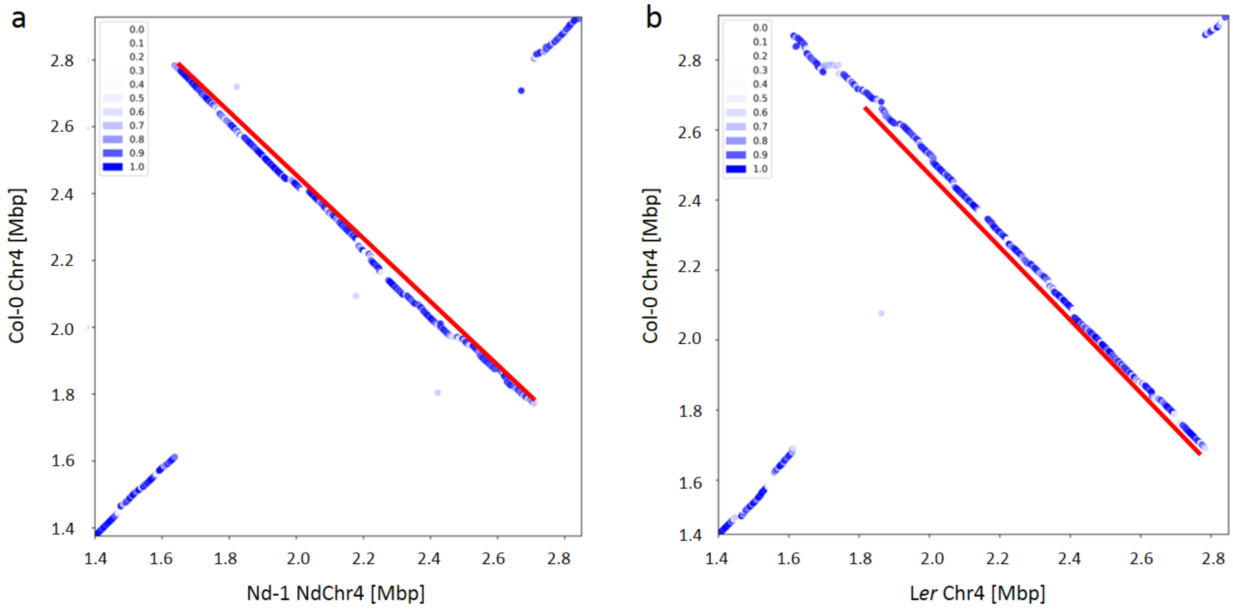

Fig.13: Inversion on chromosome 4.

The dot plot heatmaps show the similarity between small fragments of two sequences. Each dot indicates a match of 1 kbp between both sequences, while the colour is indicating the similarity of the matching sequences. Matches with low similarity are indicated by white dots, while perfect matches are indicated by blue dots. Comparison of the Nd-1 genome sequence against the Col-0 reference sequence reveals a 1 Mbp inversion (a). The L*er* genome sequence displays another inversion allele (b) [212].

2.3.3 The Nd-1 gene set

Protein encoding genes were predicted based on hints derived from the Araport11 annotation of the Col-0 reference sequence and previously optimized parameters [62]. In total, 30,126 genes with an average transcript length of 1.8 kbp were predicted. An alignment of previously generated ESTs [328] with the predicted mRNA sequences displayed an average similarity of 98.7% thus supporting the assembly and annotation quality. In total, 28,042 (93%) predicted Nd-1 genes were connected to nuclear protein encoding genes in the Araport11 annotation of Col-0 through the identification of RBHs or at least unidirectional best BLAST hits on the peptide level. After discarding TE genes from this prediction, 27,247 protein encoding genes remained. Differences between the gene sets of Nd-1 and Col-0 are usually caused by (tandem) duplications or deletions of such copies. Duplications of At2g06555 (unknown protein), At3g05530 (*RPT5A*), and At4g11510 (*RALFL28*) were validated via PCR. However, the new assembly confirms the strong overall synteny between Nd-1 and Col-0. At first, At4g22214 appeared to be another duplication example. However, one of the gene copies present in the Col-0 genome is missing in the reference sequence. Although this locus does not display a PAV or CNV between both accessions, it highlights the potential of this long read assembly to reveal mis-assemblies in the high quality reference sequence.

2.3.4 Pan-genomic analysis of *A. thaliana*

Based on the structural annotation of the Nd-1 assembly, PAVs in 964 *A. thaliana* accessions were investigated to identify dispensable genes. The coverage values of a read mapping were harnessed to infer the presence/absence of genes by comparing the average coverage across a gene to the overall average coverage value of the respective accession. In total, 1,438 genes were classified as dispensable, because these genes lacked read mapping coverage in at least 100 accessions. There are probably many more dispensable genes. However, a strict cutoff is needed to avoid a high false positive rate due to very low sequencing depth of several accessions. In agreement with previous reports [427,468], many genes associated with pathogen response were identified as dispensable. However, over 30% of the dispensable genes have no functional annotation, because no suitable match against the Araport11 annotation of Col-0 was detected or due to a lack of functional information in Araport11.

2.4 Non-canonical splice sites in plant genomes

Eukaryotic genes are split into multiple parts by introns [448]. The removal of these introns requires the binding of a spliceosome and cutting at precisely defined positions [469,470]. These splice sites are defined by highly conserved dinucleotides at the terminal intron ends: GT-AG [4]. However, the major non-canonical splice site combinations GC-AG and AT-AC are known exceptions from this rule. In addition, there were reports about other nucleotides observed at these positions which are likely to be in part artefacts of the sequencing, assembly, or annotation process [3,4]. Nevertheless, non-canonical splice sites are effecting a substantial number of genes and pose a challenge to accurate gene prediction [62,465]. A comprehensive identification of these splice site combinations is needed to understand the pattern of occurrence. A validation of annotated non-canonical splice site combinations e.g. via RNA-Seq read mapping is necessary to avoid investigating annotation artefacts and degenerated pseudogenes. In addition, this analysis of all 121 annotated plant genome sequences is intended as a resource for future studies.

2.4.1 Annotated diversity

There is a huge diversity of different minor non-canonical splice site combinations annotated throughout plant genome sequences. Even when restricting the analyses to introns within the protein encoding part of representative transcripts, there is a substantial diversity detectable. However, some clear trends emerged during the analysis (Fig.14). There is a negative correlation between the frequency of non-canonical splice site combinations and the divergence of these non-canonical splice site combinations from the canonical GT-AG combination. A strong phylogenetic signal with respect to minor non-canonical splice sites was not observed. This might be due to artefacts in the annotation or assembly. Moreover, random variations at splice sites could contribute additional noise. An advanced inspection of homologous splice site combinations could be the next step to reveal the phylogenetic history of non-canonical splice site combinations.

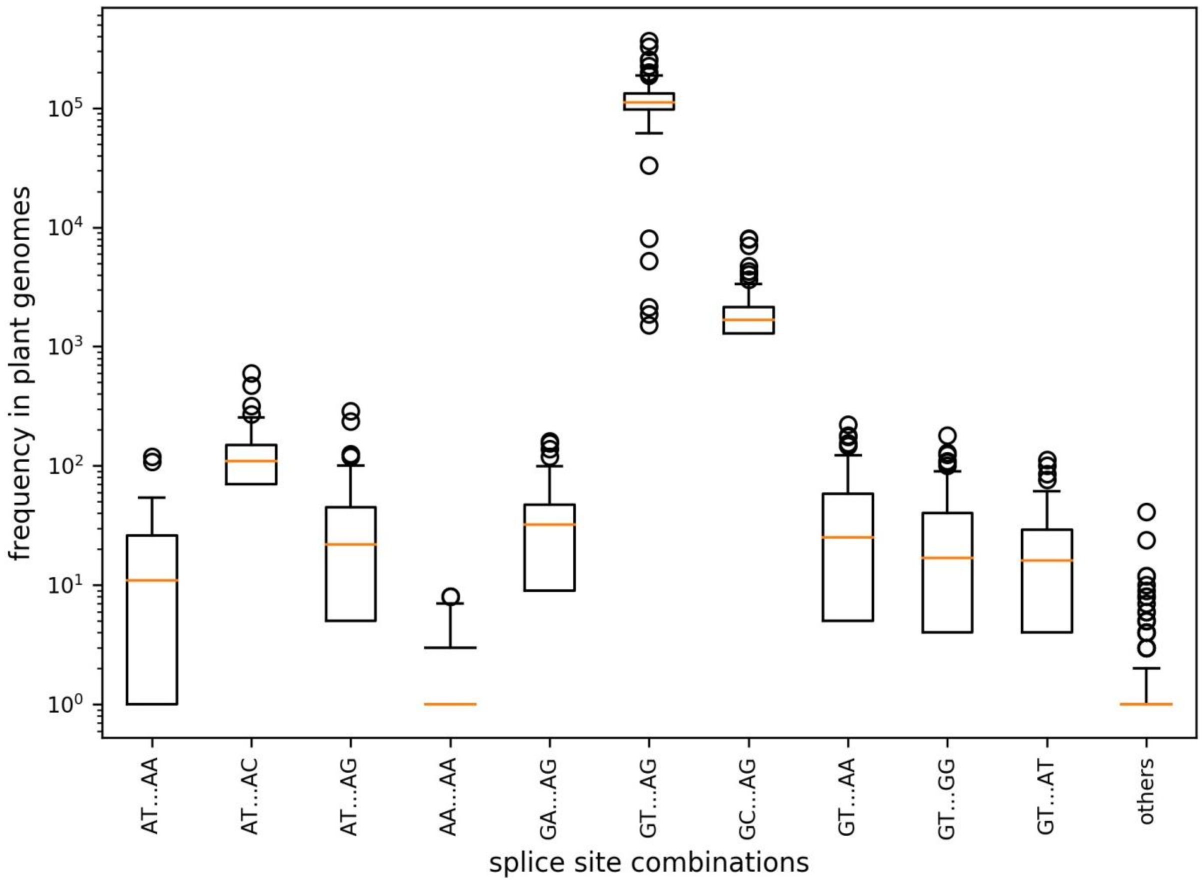

Fig.14: Splice site combination frequency.

The frequencies of selected splice site combinations across 121 plant species are displayed. Splice site combinations with high similarity to the canonical GT-AG or the major non-canonical GC-AG/AT-AC are more frequent than other splice site combinations.

2.4.2 Intron sizes in relation to splice site combinations

Comparing the sizes of introns with canonical splice site combinations and those with minor non-canonical splice site combinations revealed three major differences. First, non-canonical splice sites are more frequent in extremely short introns. Second, non-canonical splice sites are less frequent in introns with the average length of approximately 200 bp. Third, a substantially higher proportion of non-canonical introns are larger than 5 kbp.

2.4.3 Validation and usage quantification of splice sites

Splice sites were validated by assessing the coverage profiles of mapped RNA-Seq reads (Fig.15). A proper splice site that is frequently used for the splicing process would result in a high number of reads which do not contain the intron sequence. The alignment of these reads to the genome sequence results in an alignment gap over the intron. Strong differences in the coverage next to splice sites can be used to support annotations. Up to 91.3% of all annotated splice sites in representative transcripts were supported by mapped RNA-Seq reads. Donor splice sites displayed overall a stronger support than acceptor splice sites. One possible explanation is the use of alternative acceptor splice sites while there is less flexibility at the donor site. This is in agreement with previous studies which associated single donor splice sites with multiple acceptor splice sites [471,472].

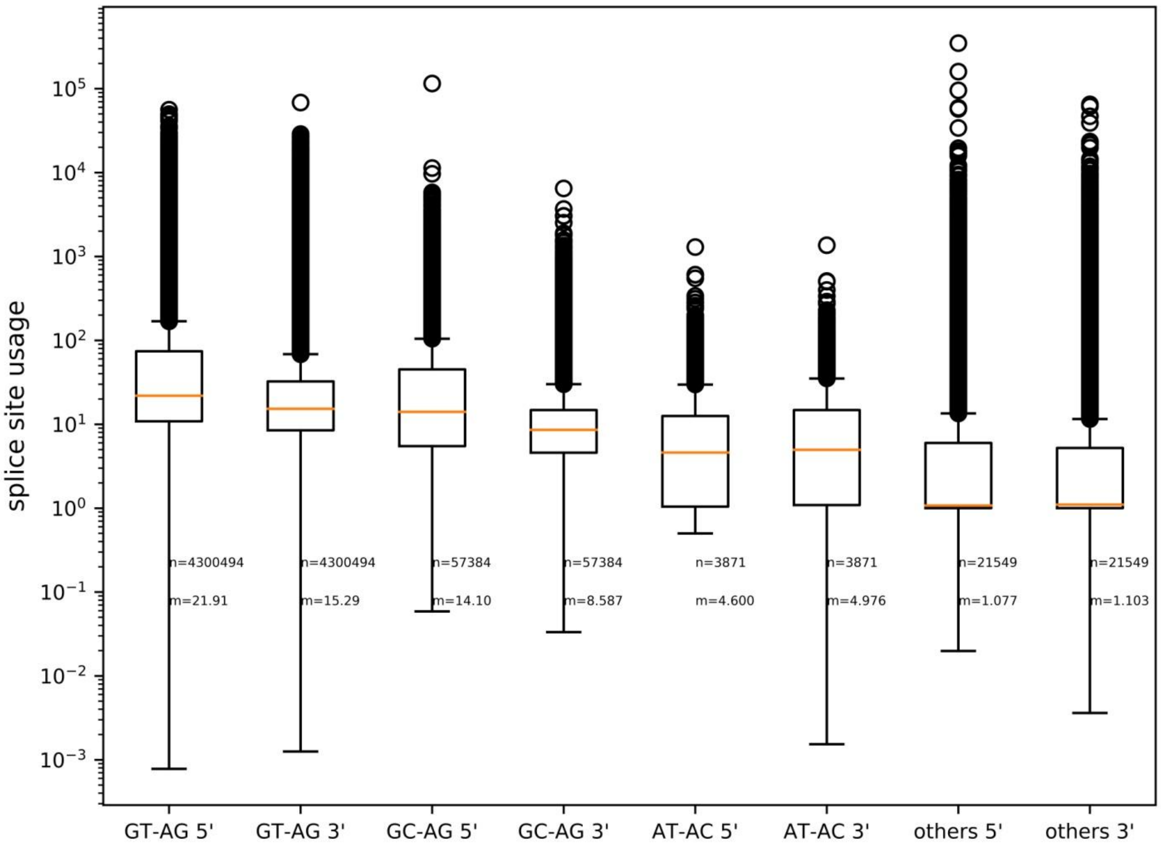

Fig.15: Usage of splice sites.

Usage of splice sites was calculated based on the number of RNA-Seq reads supporting the exon next to a splice site and the number of reads supporting the intron containing the splice site. There is a substantial difference between the usage of 5' and 3' splice sites in favour of the 5' splice sites. Canonical GT-AG splice site combinations are used more often than major or minor non-canonical splice site combinations. Sample size (n) and median (m) of the usage values are given for all splice sites.

2.4.4 Script collection for the investigation of splice site combinations

The whole investigation of non-canonical splice site combinations was performed based on dedicated Python scripts (https://github.com/bpucker/ncss2018). Initial functions were developed for the investigation of *A. thaliana* [62] and extended during the analysis of all plant genome sequences. Functions of these scripts include assessing the diversity of annotated splice site combinations, validation of these splice sites based on RNA-Seq read mappings, intron length analysis, and comparison between species. Various report files and figures are generated during the analysis process. The availability of all scripts facilitates updates of the complete analysis once more genome sequences become available. Moreover, analysing splice sites via these scripts is not restricted to the kingdom of plants, but could be applied to animals and fungi as well.

3 Discussion and outlook

This section provides an integrative discussion of the presented results and suggests directions for future analyses based on experiences collected throughout this work. Ideas and hypotheses for future studies are formulated and possible tests are suggested.

3.1 Genome sequencing and assembly

This work described two versions of an *A. thaliana* Nd-1 *de novo* genome assembly [53,213]. Although the assembly contiguity was substantially improved by long single molecule real-time (SMRT) sequencing reads, there are still genome regions missing in the second assembly. Almost 20 years after the release of the first *A. thaliana* genome sequence the currently available genome sequence is still incomplete. Centromeres and nucleolus organising regions (NORs) pose a challenge and require the routine generation of even longer reads or alternatively reads with substantially lower error rates [54,213]. There are first reports of single molecules sequenced via Oxford Nanopore Technologies (ONT) substantially exceeding the 2 Mbp mark [123]. If the read length could be further increased, this technology might have the potential to finally enable the closure of the last remaining gaps in the *A. thaliana* genome sequence. Improvements of nanopore sequencing e.g. re-reading of the very same DNA strand [75] or coupling of two nanopores with different error profiles [214] might lead to the required improvements of ONT read quality. However, latest improvements of sequencing technologies require improved DNA extraction protocols to provide high molecular input material [213,473,474]. Therefore, the bottleneck in generating even longer reads is likely to be the DNA extraction process. Efficient separation of high molecular DNA molecules from smaller fragments would be required to harness the full potential of long read sequencing technologies.

Comprehensive knowledge about the primary structure of a genome can facilitate investigations of the three-dimensional organization of DNA in the nucleus. Transcriptional regulation elements can be located far away on the sequence and are brought in physical contact with target genes through chromatin loops as observed in *A. thaliana* [475–478]. The regulation of gene expression is influenced by many factors including the chromatin structure and other epigenetic modifications [479]. Genes evolved in certain regions with local transcriptional regulation and are therefore non-randomly distributed over the genome

[480,481]. First studies investigated the three-dimensional structure of the *A. thaliana* genome via Hi-C and similar techniques [478,482,483]. In contrast to other species, topologically associated domains are not particular important in *A. thaliana* [477,482,484]. Chromatin loops in *A. thaliana* are often small and bring the 3'-end of a gene into contact with the 5'-end [478] which might facilitate transcription after initial recruitment of RNA polymerase II [477]. However, 'transcriptional factories' [485], where highly transcribed genes are clustered, were not observed in *A. thaliana* [478]. Contiguous genome sequences like the one presented in this work [213] are especially important for these analyses as a strong interaction of heterochromatic regions e.g. around the centromeres was reported as the dominant structuring force in *A. thaliana* [478,484,486]. Another Hi-C study reported an enrichment of SVs at positions with increased interchromosomal contact [487]. These regions are also assumed to display more frequently T-DNA insertions in mutagenesis experiments [487] thus indicating general susceptibility to modifications. As regions with interchromosomal proximity were reported to comprise heterochromatin e.g. repeats around the centromeres [478], this might explain the frequent observation of SVs close to the centromeres in this work [53,213].

Besides the chromatin structure, modifications of the DNA like methylation are effecting gene expression. Epigenetic investigations with focus on methylation patterns are facilitated by third generation sequencing technologies which provide the ability to analyse sequences and modifications of DNA molecules at the same time. Although ONT sequencing reads are probably better suited for the investigation of methylation patterns as more modification types can be detected, SMRT sequencing data sets with sufficient sequencing depth can be subjected to analyses of some modifications [136]. Previous studies identified methylation mechanisms in *A. thaliana* [488] and reported quantitative trait loci (QTL) for differences in the methylation patterns between different accessions [323,489]. Long reads are crucial to resolve transposable elements (TEs) and especially nested insertions [213,490]. Most TEs in euchromatic regions are highly methylated in order to suppress transposition [491–494]. Therefore, sequence and methylation information must be collected for the same molecule to investigate the regulation of TEs. In addition, the methylation of currently inaccessible regions like NORs and centromeres [54,213] could be addressed in the future.

These regions could become the target of comparative genomics with additional improvements of sequencing technologies within the next years. Since these inaccessible regions are missing in the current reference sequence TAIR9, re-sequencing experiments focusing on the rate of variant accumulation per generation [320,323,402] cannot identify variants in these regions. NORs and centromeres were previously reported to account for major genome size differences between accessions [324,415] thus high variability of these regions seems likely. Finally, improved sequencing capacities might lead to a replacement of re-sequencing projects [324,330] by *de novo* assembly and comparative genomics based on complete genome sequences [51].

When it comes to annotation of the generated genome sequences, the sequencing of full length transcripts [164,495] would be beneficial for the generation of hints. Single cell RNA-Seq [496–499] might be helpful to capture genes which are only transcribed in specific developmental stages, in rare cell types, or under certain environmental conditions. Protocols for single cell RNA-Seq of plants are still in development and will require further optimization and adaptation especially for RNA extraction from different species and cell types [498,499]. In addition, bioinformatic tools need to be adjusted to the new requirements to cope with inherent biases [499,500]. Single cell DNA sequencing could also be applied to investigate the above mentioned three-dimensional structure of the genome as cell type specific differences are expected [478].

3.2 Genome size of *A. thaliana*

Many studies investigated the genome size differences between various *A. thaliana* accessions [258,324,414,501], *Arabidopsis* species [411,422], and closely related species [409,411]. A huge proportion of these differences was previously attributed to changes in the number of 45S rDNA repeats [324,415]. This could be connected with previously observed hypomethylation of NORs and centromeres [502,503]. Investigations of a correlation between the degree of methylation in certain regions in correlation with the genome size could provide more insights. Although a contribution of additional regions and mechanisms cannot be ruled out, the relevance of NORs is supported by multiple independent studies. *A. thaliana*-derived rRNA genes are selectively silenced in *A. suecica* the allotetraploid hybrid of *A. thaliana* and *A. arenosa* [504] indicating that a loss of copies might take place without detrimental consequences. The same regulatory mechanism is expected behind the silencing of specific rRNA gene variants in *A. thaliana* during the ontogenesis [505,506]. While dimethylation of histone H3 on K9 leads to transcriptional silencing of attached rRNA genes, a trimethylation of K4 results in transcriptional activation [506–508]. Histon modifications are not just affecting transcriptional activities, but the mutation of the monomethyltransferases *ATXR5* or *ATXR6* in *A. thaliana* were also reported to cause overreplication of silenced repetitive elements [508,509]. Correlating the transcriptional activity to sequence variants and the replication efficiency could be a strategy to identify mechanisms which effect genome sizes. In summary, it can be speculated that DNA sequence variants between accessions might lead to different degrees of histon-mediated compression thus explaining differences in the replication efficiency which leads to CNVs of 45S rRNA genes ultimately resulting in genome size differences.

TEs could be another source for genome size differences between accessions. Since most TEs in *A. thaliana* appear to have no functional relevance, these elements provide material for deletions without deleterious consequences [320]. The availability of highly contiguous genome sequences of several *A. thaliana* accessions [53,54,120,211–213] enables the identification of active TEs based on comparative genomics as previously suggested [510] and could reveal novel full length elements. Available sequencing data sets of many additional accessions [53,326,327,330,511] allow the identification of PAVs of TEs through read mappings against high quality genome sequences. Due to missing regions in the reference sequence [54,213], the true proportion of TEs might be underestimated since long terminal repeats (LTRs) are expected to be abundant in the pericentromeric regions [512–514]. A comprehensive investigation of TEs in *A. thaliana* would therefore require a systematic re-annotation of a contiguous assembly according to the most recent classification system [515]. *A. thaliana* TEs appear to be smaller than *A. lyrata* TEs [516]. Thus fragmentation or deletion of repeats are likely to have contributed to the reduced genome size of *A. thaliana* compared to *A. lyrata* [422,517]. Although the reduced TE activity in selfing plants appears as a potential explanation for genome size differences, numerous small deletions in intergenic regions were previously reported to be the most important factor for the genome size difference [422]. In contrast, *Capsella rubella* was described with a genome size similar to *A. lyrata*, but differs from *A. thaliana* mainly by variants in the pericentromeric regions and probably NORs [413]. Although *C. rubella* converted to selfing like *A. thaliana*, no reduction in the amount of TEs was observed yet which could be attributed to the short phylogenetic time span [413]. An increased TE activity in *A. lyrata* compared to *A. thaliana* would be another explanation for the observed differences in TE content between both *Arabidopsis* species [394,413,422]. If TE degeneration and deletion contribute to the genome shrinkage of *A. thaliana*, control mechanisms of TE activity must differ between both *Arabidopsis* species [394,422,518]. Other hypotheses assume a reduced TE activity in *A. thaliana* which might be caused by the high gene density which causes most TE insertions to be deleterious [409]. Since numerous RNA-Seq data sets are publicly available, an investigation of TE gene transcription in *A. thaliana* is feasible. These expression values could be compared to TE gene expression in other Brassicaceae with a high number of available RNA-Seq data sets thus providing information about the relative TE transcription in *A. thaliana*.

The partial deletion of TEs is matching initial reports about a general excess of deletions over insertions in *A. thaliana* accessions [422]. These comparisons were based on chromosome sequence alignments and not on more biased short read mappings which would artificially favour deletions over insertions [102,422]. Illegitimate recombination between LTRs can cause deletions [519] and might contributed to the observed differences. High quality genome assemblies [54,212,213] could be used to validate these findings. Deletions should occur with higher allele frequencies than insertions if this mechanism is contributing to the genome shrinkage [520]. While the number of highly contiguous genome sequences is still too low for species-wide studies, the available data sets for the 1001 genomes project allow investigation of allele frequencies based on read mappings. Since recent long read assemblies are already exceeding the Col-0 reference sequence in terms of contiguity [54,213], results of new studies could surpass previous ones.

However, there must be a mechanistic explanation for previously reported differences in the genome size between accessions [258,324,414], because small differences in the genome size should not result in a selective advantage [521,522]. Differences in exonuclease activity were previously observed between *A. thaliana* and *Solanum lycopersicum* and therefore proposed as one possible explanation [523]. A recent study investigated the replication of the Col-0 genome of cells in suspension culture and classified regions of the genome into bins based on the replication in early, middle, or late S phase [524]. Correlating the replication phase of genes and regions with PAVs could be a way to find mechanistic explanations for the loss of certain regions. Genes and euchromatic regions are reported to replicate generally earlier than TEs and heterochromatic regions around the centromeres [524]. The observed enrichment of large SVs around the centromeres in this work [53,213] might be explained by errors at the end of the S phase.

Eukaryotic genomes usually contain a huge amount of introns which contribute substantially to the total genome size thus a removal of introns would contribute to genome shrinkage [404,525]. The mutational hazard hypothesis proposes the loss of introns to reduce the amount of DNA which could receive hazardous mutations [404,526]. While there is a correlation between the synonymous substitution rate and the likelihood that the effected intron is lost, selection works against the loss of introns with regulatory elements [404,527]. In general, the mutation rate in *A. thaliana* exceeds the rate in *A. lyrata* [528] thus suggesting a reason why intron loss is stronger in *A. thaliana*.

In summary, different factors are likely to shape the *A. thaliana* genome: i) differences in 45S rRNA gene replication due to methylation differences, ii) differences in TE activity, iii), favouring deletions over insertions, and iv) loss of introns.

3.3 Gene set differences

High quality genome sequences of several *A. thaliana* accessions [53,54,120,211–213] and sequencing data sets of many additional accessions [330] allow a detailed pan-genome analysis e.g. the classification of genes as ‘essential’ or ‘dispensable’. However, the situation is slightly more complex with intermediate genes which could be labelled ‘conditionally dispensable’ [246]. Since an organism with a truly minimal gene set would require an environment without selection pressure, the minimal gene set is a theoretical concept which is unlikely to be reached [529].

Previous re-sequencing studies reported about 300-500 presence/absence variants (PAVs) of genes per *A. thaliana* accession [320,427]. The number of apparently dispensable genes increases when additional accessions are included in the analysis as the investigation of 1,135 accessions demonstrated by identifying 17,692 genes with at least one high impact variant [330]. Despite numerous false positives, this number might still be an underestimation due to the remaining gaps and errors in the Col-0 reference sequence which was used for the mapping-based identification of functionless genes. Due to at least three duplications of the genome of an *A. thaliana* ancestor, redundant gene copies can be removed without phenotypic impact. It will be interesting to see in the future if the incorporation of more sequencing data sets derived from additional accessions will lead to a saturation of observed absent genes. In this work, copy number variations (CNVs) and

PAVs were also identified based on the coverage in read mappings [53,213]. The number of gene differences between Col-0 and Nd-1 is in the same range as previous reports [320,427]. This method is probably still the best approach to perform pan-genomic investigations in *A. thaliana*. The risk of missing genes in the selected reference sequence decreases with increasing assembly quality. However, variations in the phylogenetic distances of the analysed accessions could bias the results due to differences in the mappability of reads. Another issue are differences in the applied sequencing technologies and therefore different coverage biases.

Most previous studies were only focused on differences concerning protein encoding genes, but high quality sequences combined with advanced RNA-Seq workflows facilitate the investigation of various RNA genes. The annotation of non-coding RNA genes can be performed via Rfam [530] and other tools [531,532]. miRBase [533,534] is frequently applied for the annotation of miRNA genes and the prediction of tRNA genes is often based on tRNAscan-SE [535]. These tools provide the basis to expand comparative genomics to the RNA gene level. Since the focus of Araport11 was on the annotation of non-coding RNAs by incorporating information from RNA-Seq experiments, a comprehensive analysis of RNA genes across numerous *A. thaliana* accessions could add a pan-genomic perspective.

3.4 Non-canonical splice sites

After analysing the presence of non-canonical splice site combinations on a massive scale in plants [7,62], the functional impact of these splice site combinations needs to be assessed in much greater detail. Several suggestions for further experiments and current limitations that need to be addressed were already described [7]. This chapter illustrates some opportunities for future research including the analysis of i) an extended data set, ii) the impact of environmental factors on non-canonical splice sites, and iii) approaches to overcome technical limitations.

A comparison of the non-canonical splice site combinations detected in plants to splice site combinations in animals and fungi would be interesting to identify plant-specific characteristics and mechanisms. Therefore, the established analyses [7] can be applied to an extended collection of data sets by re-using the collection of Python scripts. Since high contiguity genome sequences become available with increasing pace, it might be interesting to repeat such analyses in the near future to harness the power of substantially larger data sets. Although no phylogenetic signal of non-canonical splice site combinations was observed yet [7], a pattern might emerge when more closely related species are included.

In general, alternative splicing is substantially influenced by external factors e.g. light intensities, salt concentrations, or pathogens [536–538]. A comprehensive investigation of the usage of non-canonical splice site combinations under various conditions might reveal the relevance of these splice sites and could even facilitate the detection of molecular mechanisms involved in the splicing. As already speculated, certain splice site combinations might be associated with very precisely defined conditions [7]. It is important to distinguish between a potential dedication of non-canonical splice site combinations to specific environmental conditions and the expression of different genes, which happen to display such non-canonical splice sites. Performing the same analysis for multiple species could help to distinguish both possibilities. Another dimension is ontogenesis. Numerous reports describe changes in splicing patterns during the development and in different cell types [539–541]. Again, it would be interesting to see if certain non-canonical splice site combinations are dedicated to certain developmental stages.

If non-canonical splice sites are used, what is the proportion of resulting transcripts compared to all transcripts of the gene? Relative quantification of different transcript isoforms could help to shed more light on the evolution of non-canonical splice sites. Especially changes herein over evolutionary times would be interesting. While isolated splice sites can be analysed based on common RNA-Seq data sets, the precise quantification of different transcript isoforms is still challenging. These analyses could be facilitated by high-throughput long read sequencing technologies to recover complete transcript sequences in a single read.

On the technical level, the comparison of different annotations e.g. NCBI and Araport11 [2] for *A. thaliana* could reveal insights into workflow specific differences in the annotation of non-canonical splice site combinations. Similar comparisons are possible for many organisms for which the community is curating several annotations in parallel e.g. *Vitis vinifera*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, and *Homo sapiens*. Different annotation tools and approaches were applied to data sets from different sources to generate these annotations. Systematic differences might exist between popular annotation collections provided by the NCBI [542] and phytozome [543]. Results of these comparisons could help to further improve gene prediction workflows.

The investigation of splice site usage is relying on the accurate placement of RNA-Seq reads to a genomic reference sequence which requires an alignment with gaps at the intron positions. There are several tools dedicated to this purpose, which might show differences in terms of performance. Assessing the best available and frequently applied tools like STAR [158,286], HiSat2 [159], TopHat2 [544], and exonerate [289] could further facilitate research on splice sites and might reveal explanations for currently observed annotation differences [7]. Such a benchmarking study should not just identify the most suitable tool for a specific application, but also provide insights into the best choice of parameters. The detection of novel non-canonical splice site combinations based on RNA-Seq data sets would benefit from these benchmarking results. This step from the assessment of existing annotations to the identification of novel non-canonical splice sites would substantially increase the accessible taxonomic diversity as genome sequences without annotation could be included in the analysis.

3.5 Transfer to crops

The investigation of non-canonical splice sites already demonstrated how methods and knowledge can be transferred from the model organism *A. thaliana* to other plant species including crops. Not just the assembled genome sequence [1,384], but also the high quality annotation [2,384] is crucial for plant biotechnology. Investigations of the (genome) evolution of *A. thaliana* revealed insights [23] which are also improving our understanding of the phylogenetic development of other plant species. A huge diversity of *A. thaliana* accessions allows the analysis of local adaptations as this inbreeding species preserves genes which originated in specific geographic regions [545]. Based on citations, *A. thaliana* research is even more important for research on other species than for research on this model organism itself [546].

Information is transferred from *A. thaliana* to other plant species on various levels including sequences and parameters for gene prediction [21,112,219,413,422,547] as well as functional annotations [24,113,363,548]. Moreover, *A. thaliana* is usually included in comparative genomics [21,112,219,413,422,547,549] and phylogenetic analyses

[11,21,230,413]. Thus, an improved understanding of the *A. thaliana* pan-genome and an improved reference sequence will result in enhanced annotations of other genome sequences.

The Nd-1 genome sequence and all related genomic resources provide the basis for benchmarking studies. Different second generation sequencing technologies and SMRT sequencing were applied to analyse the same biological material. Since the detection of sequence variants is frequently required in re-sequencing projects or in mapping-by-sequencing studies [32,550], the Illumina sequencing reads could be used to optimize variant calling parameters in read mappings against the Col-0 reference sequence. The availability of a highly contiguous Nd-1 genome sequence based on independent SMRT sequencing reads provides the opportunity to validate sequence variants *in silico*. To the best of my knowledge, there is no benchmarking study about the best tools and parameters for variant calling in plants.

The consideration of non-canonical splice sites during gene prediction on new crop genome sequences can lead to a higher quality of the structural annotation. One example with a non-canonical splice site combination in the gene structure is *FGT1* (At1g79350) which was investigated in *A. thaliana* Col-0 and Nd-1 [7,62]. Homologous genes with non-canonical splice sites were discovered in other Brassicaceae including the crops *Brassica napus*, *B. oleracea*, and *B. rapa* [7]. Since *FGT1* was reported to mediate chromatin memory in response to stress [551], a proper structural annotation is beneficial to understand stress responses in these crop species. Breeding programs might benefit from an accurate gene structure, because non-functional alleles were observed to cause too early down-regulation of stress response genes after heat exposure [551]. *FGT1* is only one example for a gene with non-canonical splice sites in *A. thaliana* which could facilitate crop improvements if the knowledge is transferred. A systematic investigation of all genes with non-canonical splice site combinations is likely to reveal more cases.

3.6 Conclusion

In summary, this work contributes to the research on *A. thaliana* by generating one of the most contiguous genome assemblies for this species together with an optimized structural annotation. This genomic resource revealed numerous variants and will facilitate *A. thaliana* pan-genomics in the future. The investigation of non-canonical splice sites in this model organism paved the way for an extended study across the kingdom of plants. This is a successful example how methods and knowledge gained from research on *A. thaliana* were transferred to other plants including important crop species. The extended analysis revealed insights about the impact of non-canonical splice sites and provides a comprehensive resource for future studies.

4 References

1. *Arabidopsis* Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408: 796–815. doi:10.1038/35048692
2. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. Plant J. 2017;89: 789–804. doi:10.1111/tpj.13415
3. Jackson IJ. A reappraisal of non-consensus mRNA splice sites. Nucleic Acids Res. 1991;19: 3795–3798.
4. Burset M, Seledtsov IA, Solovyev VV. Analysis of canonical and non-canonical splice sites in mammalian genomes. Nucleic Acids Res. 2000;28: 4364–4375.
5. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3—new capabilities and interfaces. Nucleic Acids Res. 2012;40: e115. doi:10.1093/nar/gks596
6. Guo D, Song X, Yuan M, Wang Z, Ge W, Wang L, et al. RNA-Seq Profiling Shows Divergent Gene Expression Patterns in *Arabidopsis* Grown under Different Densities. Front Plant Sci. 2017;8. doi:10.3389/fpls.2017.02001
7. Pucker B, Brockington SF. Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes. BMC Genomics. 2018;19: 980. doi:10.1186/s12864-018-5360-z
8. Nakahama K, Urata N, Shinya T, Hayashi K, Nanto K, Rosa AC, et al. RNA-seq analysis of lignocellulose-related genes in hybrid *Eucalyptus* with contrasting wood basic density. BMC Plant Biol. 2018;18: 156. doi:10.1186/s12870-018-1371-9
9. Robinson AJ, Tamiru M, Salby R, Bolitho C, Williams A, Huggard S, et al. AgriSeqDB: an online RNA-Seq database for functional studies of agriculturally relevant plant species. BMC Plant Biol. 2018;18: 200. doi:10.1186/s12870-018-1406-2
10. Shiu S-H, Shih M-C, Li W-H. Transcription Factor Families Have Much Higher Expansion Rates in Plants than in Animals. Plant Physiol. 2005;139: 18–26. doi:10.1104/pp.105.065110
11. Stracke R, Holtgräwe D, Schneider J, Pucker B, Sörensen TR, Weisshaar B. Genome-wide identification and characterisation of R2R3-MYB genes in sugar beet (*Beta vulgaris*). BMC Plant Biol. 2014;14: 249. doi:10.1186/s12870-014-0249-8
12. Xie T, Chen C, Li C, Liu J, Liu C, He Y. Genome-wide investigation of *WRKY* gene family in pineapple: evolution and expression profiles during development and stress. BMC Genomics. 2018;19. doi:10.1186/s12864-018-4880-x
13. Pinosio S, Giacomello S, Faivre-Rampant P, Taylor G, Jorge V, Le Paslier MC, et al. Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. Mol Biol Evol. 2016;33: 2706–2719. doi:10.1093/molbev/msw161
14. Zhang Z, Mao L, Chen H, Bu F, Li G, Sun J, et al. Genome-Wide Mapping of Structural Variations Reveals a Copy Number Variant That Determines Reproductive Morphology in Cucumber. Plant Cell. 2015;27: 1595–1604. doi:10.1105/tpc.114.135848
15. Saxena RK, Edwards D, Varshney RK. Structural variations in plant genomes. Brief Funct Genomics. 2014;13: 296–307. doi:10.1093/bfgp/elu016
16. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes. 2007;7: 574–578. doi:10.1111/j.1471-8286.2007.01758.x
17. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17: 81–92. doi:10.1038/nrg.2015.28
18. Ellegren H, Galtier N. Determinants of genetic diversity. Nat Rev Genet. 2016;17: 422–433. doi:10.1038/nrg.2016.58
19. Jordan KW, Wang S, Lun Y, Gardiner L-J, MacLachlan R, Hucl P, et al. A haplotype map of allohexaploid wheat reveals distinct patterns of selection on homoeologous genomes. Genome Biol. 2015;16: 48. doi:10.1186/s13059-015-0606-4
20. Macko-Podgórni A, Machaj G, Stelmach K, Senalik D, Grzebelus E, Iorizzo M, et al. Characterization of a Genomic Region under Selection in Cultivated Carrot (*Daucus carota* subsp. *sativus*) Reveals a Candidate Domestication Gene. Front Plant Sci. 2017;8: 12. doi:10.3389/fpls.2017.00012
21. Dohm JC, Minoche AE, Holtgräwe D, Capella-Gutiérrez S, Zakrzewski F, Tafer H, et al. The genome of the recently domesticated crop plant sugar beet (*Beta vulgaris*). Nature. 2014;505: 546–549. doi:10.1038/nature12817
22. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of Gene Duplication in Plants. Plant Physiol. 2016;171: 2294–2316. doi:10.1104/pp.16.00523
23. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome Biol. 2016;17: 37. doi:10.1186/s13059-016-0908-1
24. Badouin H, Gouzy J, Grassa CJ, Murat F, Staton SE, Cottret L, et al. The sunflower genome provides insights into oil metabolism, flowering and Asterid evolution. Nature. 2017;546: 148–152. doi:10.1038/nature22380
25. Sun G, Xu Y, Liu H, Sun T, Zhang J, Hettenhausen C, et al. Large-scale gene losses underlie the genome evolution of parasitic plant *Cuscuta australis*. Nat Commun. 2018;9: 2683. doi:10.1038/s41467-018-04721-8
26. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010;42: 961–967. doi:10.1038/ng.695
27. Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet. 2012;44: 32–39. doi:10.1038/ng.1018
28. Fechter I, Hausmann L, Zyprian E, Daum M, Holtgräwe D, Weisshaar B, et al. QTL analysis of flowering time and ripening traits suggests an impact of a genomic region on linkage group 1 in *Vitis*. TAG Theor Appl Genet Theor Angew Genet. 2014;127: 1857–1872. doi:10.1007/s00122-014-2310-2
29. Collard BCY, Jahufer MZZ, Brouwer JB, Pang ECK. An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica. 2005;142: 169–196. doi:10.1007/s10681-005-1681-5
30. Varshney RK, Graner A, Sorrells ME. Genomics-assisted breeding for crop improvement. Trends Plant Sci. 2005;10: 621–630. doi:10.1016/j.tplants.2005.10.004

31. Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, et al. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. Proc Natl Acad Sci. 2010;107: 10578–10583. doi:10.1073/pnas.1005931107
32. He J, Zhao X, Laroche A, Lu Z-X, Liu H, Li Z. Genotyping-by-sequencing (GBS), an ultimate marker-assisted selection (MAS) tool to accelerate plant breeding. Front Plant Sci. 2014;5. doi:10.3389/fpls.2014.00484
33. Barabaschi D, Tondelli A, Desiderio F, Volante A, Vaccino P, Valè G, et al. Next generation breeding. Plant Sci Int J Exp Plant Biol. 2016;242: 3–13. doi:10.1016/j.plantsci.2015.07.010
34. Bevan MW, Uauy C, Wulff BBH, Zhou J, Krasileva K, Clark MD. Genomic innovation for crop improvement. Nature. 2017;543: 346–354. doi:10.1038/nature22011
35. Liu H, Ding Y, Zhou Y, Jin W, Xie K, Chen L-L. CRISPR-P 2.0: An Improved CRISPR-Cas9 Tool for Genome Editing in Plants. Mol Plant. 2017;10: 530–532. doi:10.1016/j.molp.2017.01.003
36. Rodríguez-Leal D, Lemmon ZH, Man J, Bartlett ME, Lippman ZB. Engineering Quantitative Trait Variation for Crop Improvement by Genome Editing. Cell. 2017;171: 470–480.e8. doi:10.1016/j.cell.2017.08.030
37. Spindel JE, McCouch SR. When more is better: how data sharing would accelerate genomic selection of crop plants. New Phytol. 2016;212: 814–826. doi:10.1111/nph.14174
38. Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J. Breeding schemes for the implementation of genomic selection in wheat (*Triticum* spp.). Plant Sci. 2016;242: 23–36. doi:10.1016/j.plantsci.2015.08.021
39. Vivek BS, Krishna GK, Vengadessan V, Babu R, Zaidi PH, Kha LQ, et al. Use of Genomic Estimated Breeding Values Results in Rapid Genetic Gains for Drought Tolerance in Maize. Plant Genome. 2017;10. doi:10.3835/plantgenome2016.07.0070
40. Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 2009;27: 522–530. doi:10.1016/j.tibtech.2009.05.006
41. Massawe F, Mayes S, Cheng A. Crop Diversity: An Unexploited Treasure Trove for Food Security. Trends Plant Sci. 2016;21: 365–368. doi:10.1016/j.tplants.2016.02.006
42. Chang Y, Liu H, Liu M, Liao X, Sahu SK, Fu Y, et al. The draft genomes of five agriculturally important African orphan crops. GigaScience. 2018; doi:10.1093/gigascience/giy152
43. Varshney RK, Close TJ, Singh NK, Hoisington DA, Cook DR. Orphan legume crops enter the genomics era! Curr Opin Plant Biol. 2009;12: 202–210. doi:10.1016/j.pbi.2008.12.004
44. McCouch SR, Sweeney M, Li J, Jiang H, Thomson M, Septiningsih E, et al. Through the genetic bottleneck: *O. rufipogon* as a source of trait-enhancing alleles for *O. sativa*. Euphytica. 2007;154: 317–339. doi:10.1007/s10681-006-9210-8
45. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol. 2012;30: 105–111. doi:10.1038/nbt.2050
46. Munns R, James RA, Xu B, Athman A, Conn SJ, Jordans C, et al. Wheat grain yield on saline soils is improved by an ancestral Na^+ transporter gene. Nat Biotechnol. 2012;30: 360–364. doi:10.1038/nbt.2120
47. Zhang H, Mittal N, Leamy LJ, Barazani O, Song B-H. Back into the wild—Apply untapped genetic diversity of wild relatives for crop improvement. Evol Appl. 2017;10: 5–24. doi:10.1111/eva.12434
48. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Res. 2016;44: e89. doi:10.1093/nar/gkw092
49. Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC Bioinformatics. 2018;19: 189. doi:10.1186/s12859-018-2203-5
50. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinforma Oxf Engl. 2014;30: 2114–2120. doi:10.1093/bioinformatics/btu170
51. Jiao W-B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017;36: 64–70. doi:10.1016/j.pbi.2017.02.002
52. Chen F, Dong W, Zhang J, Guo X, Chen J, Wang Z, et al. The Sequenced Angiosperm Genomes and Genome Databases. Front Plant Sci. 2018;9. doi:10.3389/fpls.2018.00418
53. Pucker B, Holtgräwe D, Sörensen TR, Stracke R, Viehöver P, Weisshaar B. A *De Novo* Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny. PLOS ONE. 2016;11: e0164321. doi:10.1371/journal.pone.0164321
54. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity *Arabidopsis thaliana* genome assembly with a single nanopore flow cell. Nat Commun. 2018;9: 541. doi:10.1038/s41467-018-03016-2
55. Wetterstrand K. DNA Sequencing Costs: Data: Data from the NHGRI Genome Sequencing Program (GSP). In: National Human Genome Research Institute (NHGRI) [Internet]. 2018 [cited 5 Jan 2019]. Available: https://www.genome.gov/sequencingcostsdata
56. List of sequenced plant genomes [Internet]. Wikipedia. 2019. Available: https://en.wikipedia.org/w/index.php?title=List_of_sequenced_plant_genomes&oldid=876968751
57. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94: 441–448.
58. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74: 5463–5467.
59. Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74: 560–564.
60. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010;11: 31–46. doi:10.1038/nrg2626
61. Beck TF, Mullikin JC, Biesecker LG. Systematic Evaluation of Sanger Validation of NextGen Sequencing Variants. Clin Chem. 2016;62: 647–654. doi:10.1373/clinchem.2015.249623
62. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. BMC Res Notes. 2017;10. doi:https://doi.org/10.1186/s13104-017-2985-y
63. Zong Y, Wang Y, Li C, Zhang R, Chen K, Ran Y, et al. Precise base editing in rice, wheat and maize with a Cas9-cytidine deaminase fusion. Nat Biotechnol. 2017;35: 438–440. doi:10.1038/nbt.3811

64. Liang Z, Chen K, Li T, Zhang Y, Wang Y, Zhao Q, et al. Efficient DNA-free genome editing of bread wheat using CRISPR/Cas9 ribonucleoprotein complexes. Nat Commun. 2017;8: 14261. doi:10.1038/ncomms14261
65. Nishihara M, Higuchi A, Watanabe A, Tasaki K. Application of the CRISPR/Cas9 system for modification of flower color in *Torenia fournieri*. BMC Plant Biol. 2018;18: 331. doi:10.1186/s12870-018-1539-3
66. Hebert PDN, Braukmann TWA, Prosser SWJ, Ratnasingham S, deWaard JR, Ivanova NV, et al. A Sequel to Sanger: amplicon sequencing that scales. BMC Genomics. 2018;19: 219. doi:10.1186/s12864-018-4611-3
67. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nat Rev Genet. 2004;5: 335–344. doi:10.1038/nrg1325
68. Shendure J, Ji H. Next-generation DNA sequencing. Nat Biotechnol. 2008;26: 1135–1145. doi:10.1038/nbt1486
69. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19: R227–R240. doi:10.1093/hmg/ddq416
70. Glenn TC. Field guide to next-generation DNA sequencers. Mol Ecol Resour. 2011;11: 759–769. doi:10.1111/j.1755-0998.2011.03024.x
71. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13: 341. doi:10.1186/1471-2164-13-341
72. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17: 333–351. doi:10.1038/nrg.2016.49
73. Ye C, Hill CM, Wu S, Ruan J, Ma Z (Sam). DBG2OLC: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. Sci Rep. 2016;6. doi:10.1038/srep31900
74. Peterson DG, Arick M. Sequencing Plant Genomes. 2018; 1–85. doi:10.1007/124_2018_18
75. Deamer D, Akeson M, Branton D. Three decades of nanopore sequencing. Nat Biotechnol. 2016;34: 518–524. doi:10.1038/nbt.3423
76. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323: 133–138. doi:10.1126/science.1162986
77. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10: 57–63. doi:10.1038/nrg2484
78. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437: 376–380. doi:10.1038/nature03959
79. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;456: 53–59. doi:10.1038/nature07517
80. Pandey V, Nutter RC, Prediger E. Applied Biosystems SOLiD^{TM} System: Ligation-Based Sequencing. Next Generation Genome Sequencing. John Wiley & Sons, Ltd; 2008. pp. 29–42. doi:10.1002/9783527625130.ch3
81. Mardis ER. The impact of next-generation sequencing technology on genetics. Trends Genet TIG. 2008;24: 133–141. doi:10.1016/j.tig.2007.12.007
82. Mardis ER. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet. 2008;9: 387–402. doi:10.1146/annurev.genom.9.081307.164359
83. Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive assembly of pyrosequencing reads with mates. Bioinforma Oxf Engl. 2008;24: 2818–2824. doi:10.1093/bioinformatics/btn548
84. Rothberg JM, Hinz W, Rearick TM, Schultz J, Mileski W, Davey M, et al. An integrated semiconductor device enabling non-optical genome sequencing. Nature. 2011;475: 348–352. doi:10.1038/nature10242
85. Ju J, Kim DH, Bi L, Meng Q, Bai X, Li Z, et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. Proc Natl Acad Sci. 2006;103: 19635–19640. doi:10.1073/pnas.0609513103
86. Guo J, Xu N, Li Z, Zhang S, Wu J, Kim DH, et al. Four-color DNA sequencing with 3’-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. Proc Natl Acad Sci U S A. 2008;105: 9145–9150. doi:10.1073/pnas.0804023105
87. Rieber N, Zapatka M, Lasitschka B, Jones D, Northcott P, Hutter B, et al. Coverage Bias and Sensitivity of Variant Calling for Four Whole-genome Sequencing Technologies. PLoS ONE. 2013;8. doi:10.1371/journal.pone.0066621
88. Dohm JC, Lottaz C, Borodina T, Himmelbauer H. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. 2008;36: e105. doi:10.1093/nar/gkn425
89. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, et al. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. 2011;39: e90. doi:10.1093/nar/gkr344
90. Hillier LW, Marth GT, Quinlan AR, Dooling D, Fewell G, Barnett D, et al. Whole-genome sequencing and variant discovery in *C. elegans*. Nat Methods. 2008;5: 183–188. doi:10.1038/nmeth.1179
91. Harismendy O, Ng PC, Strausberg RL, Wang X, Stockwell TB, Beeson KY, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. Genome Biol. 2009;10: R32. doi:10.1186/gb-2009-10-3-r32
92. O’Connell J, Schulz-Trieglaff O, Carlson E, Hims MM, Gormley NA, Cox AJ. NxTrim: optimized trimming of Illumina mate pair reads. Bioinformatics. 2015;31: 2035–2037. doi:10.1093/bioinformatics/btv057
93. Lieberman-Aiden E, Berkum NL van, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science. 2009;326: 289–293. doi:10.1126/science.1181369
94. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31: 1119–1125. doi:10.1038/nbt.2727
95. Selvaraj S, Dixon JR, Bansal V, Ren B. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. Nat Biotechnol. 2013;31: 1111–1118. doi:10.1038/nbt.2728
96. Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239: 487–491.
97. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, et al. A large genome centre’s improvements to the Illumina sequencing system. Nat Methods. 2008;5: 1005–1010. doi:10.1038/nmeth.1270

98. Kozarewa I, Ning Z, Quail MA, Sanders MJ, Berriman M, Turner DJ. Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of GC-biased genomes. Nat Methods. 2009;6: 291–295. doi:10.1038/nmeth.1311

99. Stein LD. The case for cloud computing in genome informatics. Genome Biol. 2010;11: 207. doi:10.1186/gb-2010-11-5-207

100. Hudson ME. Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour. 2008;8: 3–17. doi:10.1111/j.1471-8286.2007.02019.x

101. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, et al. The *Sorghum bicolor* genome and the diversification of grasses. Nature. 2009;457: 551–556. doi:10.1038/nature07723

102. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. Nat Methods. 2011;8: 61–65. doi:10.1038/nmeth.1527

103. Characterization TF-IPC for GG, Jaillon O, Aury J-M, Noel B, Policriti A, Clepet C, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449: 463–467. doi:10.1038/nature06148

104. International Barley Genome Sequencing Consortium, Mayer KFX, Waugh R, Brown JWS, Schulman A, Langridge P, et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491: 711–716. doi:10.1038/nature11543

105. Visendi P, Berkman PJ, Hayashi S, Golicz AA, Bayer PE, Ruperao P, et al. An efficient approach to BAC based assembly of complex genomes. Plant Methods. 2016;12: 2. doi:10.1186/s13007-016-0107-9

106. Okura VK, de Souza RSC, de Siqueira Tada SF, Arruda P. BAC-Pool Sequencing and Assembly of 19 Mb of the Complex Sugarcane Genome. Front Plant Sci. 2016;7: 342. doi:10.3389/fpls.2016.00342

107. Koganebuchi K, Gakuhari T, Takeshima H, Sato K, Fujii K, Kumabe T, et al. A new targeted capture method using bacterial artificial chromosome (BAC) libraries as baits for sequencing relatively large genes. PLOS ONE. 2018;13: e0200170. doi:10.1371/journal.pone.0200170

108. Borzęcka E, Hawliczek-Strulak A, Bolibok L, Gawroński P, Tofil K, Milczarski P, et al. Effective BAC clone anchoring with genotyping-by-sequencing and Diversity Arrays Technology in a large genome cereal rye. Sci Rep. 2018;8: 8428–8428. doi:10.1038/s41598-018-26541-y

109. Lux M, Krüger J, Rinke C, Maus I, Schlüter A, Woyke T, et al. acdc – Automated Contamination Detection and Confidence estimation for single-cell genome data. BMC Bioinformatics. 2016;17. doi:10.1186/s12859-016-1397-7

110. Tennessen K, Andersen E, Clingenpeel S, Rinke C, Lundberg DS, Han J, et al. ProDeGe: a computational protocol for fully automated decontamination of genomes. ISME J. 2016;10: 269–272. doi:10.1038/ismej.2015.100

111. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods San Diego Calif. 2013;63: 41–49. doi:10.1016/j.ymeth.2013.06.027

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

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

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

115. Branton D, Deamer DW, Marziali A, Bayley H, Benner SA, Butler T, et al. The potential and challenges of nanopore sequencing. Nat Biotechnol. 2008;26: 1146–1153. doi:10.1038/nbt.1495

116. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, Webb WW. Zero-mode waveguides for single-molecule analysis at high concentrations. Science. 2003;299: 682–686. doi:10.1126/science.1079700

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

118. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012;13: 375. doi:10.1186/1471-2164-13-375

119. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, et al. Reducing assembly complexity of microbial genomes with single-molecule sequencing. Genome Biol. 2013;14: R101. doi:10.1186/gb-2013-14-9-r101

120. Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, et al. Long-read, whole-genome shotgun sequence data for five model organisms. Sci Data. 2014;1. doi:10.1038/sdata.2014.45

121. VanBuren R, Bryant D, Edger PP, Tang H, Burgess D, Challabathula D, et al. Single-molecule sequencing of the desiccation-tolerant grass *Oropetium thomaeum*. Nature. 2015;527: 508–511. doi:10.1038/nature15714

122. Quick J, Quinlan AR, Loman NJ. A reference bacterial genome dataset generated on the MinION^{TM} portable single-molecule nanopore sequencer. GigaScience. 2014;3: 22. doi:10.1186/2047-217X-3-22

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

124. Butler TZ, Pavlenok M, Derrington IM, Niederweis M, Gundlach JH. Single-molecule DNA detection with an engineered MspA protein nanopore. Proc Natl Acad Sci U S A. 2008;105: 20647–20652. doi:10.1073/pnas.0807514106

125. Clarke J, Wu H-C, Jayasinghe L, Patel A, Reid S, Bayley H. Continuous base identification for single-molecule nanopore DNA sequencing. Nat Nanotechnol. 2009;4: 265–270. doi:10.1038/nnano.2009.12

126. Laszlo AH, Derrington IM, Ross BC, Brinkerhoff H, Adey A, Nova IC, et al. Decoding long nanopore sequencing reads of natural DNA. Nat Biotechnol. 2014;32: 829–833. doi:10.1038/nbt.2950

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

128. Derrington IM, Butler TZ, Collins MD, Manrao E, Pavlenok M, Niederweis M, et al. Nanopore DNA sequencing with MspA. Proc Natl Acad Sci. 2010;107: 16060–16065. doi:10.1073/pnas.1001831107

References

129. Manrao EA, Derrington IM, Langford KW, Pavlenok M, Niederweis M, Gundlach JH. Biological Nanopore MspA for DNA Sequencing. Biophys J. 2012;102: 203a. doi:10.1016/j.bpj.2011.11.1105
130. Smith AM, Abu-Shumays R, Akeson M, Bernick DL. Capture, Unfolding, and Detection of Individual tRNA Molecules Using a Nanopore Device. Front Bioeng Biotechnol. 2015;3: 91. doi:10.3389/fbioe.2015.00091
131. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. Nat Methods. 2018;15: 201–206. doi:10.1038/nmeth.4577
132. Nivala J, Marks DB, Akeson M. Unfoldase-mediated protein translocation through an α-hemolysin nanopore. Nat Biotechnol. 2013;31: 247–250. doi:10.1038/nbt.2503
133. Rang FJ, Kloosterman WP, de Ridder J. From squiggle to basepair: computational approaches for improving nanopore sequencing read accuracy. Genome Biol. 2018;19: 90. doi:10.1186/s13059-018-1462-9
134. Schreiber J, Wescoe ZL, Abu-Shumays R, Vivian JT, Baatar B, Karplus K, et al. Error rates for nanopore discrimination among cytosine, methylcytosine, and hydroxymethylcytosine along individual DNA strands. Proc Natl Acad Sci U S A. 2013;110: 18910–18915. doi:10.1073/pnas.1310615110
135. Karsten L, Bergen D, Drake C, Dymek S, Edich M, Haak M, et al. Expanding The Genetic Code. 2017. doi:https://doi.org/10.13140/RG.2.2.20342.91203
136. Stoiber MH, Quick J, Egan R, Lee JE, Celniker SE, Neely R, et al. *De novo* Identification of DNA Modifications Enabled by Genome-Guided Nanopore Signal Processing. bioRxiv. 2017; 094672. doi:10.1101/094672
137. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14: 407–410. doi:10.1038/nmeth.4184
138. Rand AC, Jain M, Eizenga JM, Musselman-Brown A, Olsen HE, Akeson M, et al. Mapping DNA Methylation with High Throughput Nanopore Sequencing. Nat Methods. 2017;14: 411–413. doi:10.1038/nmeth.4189
139. Ip CLC, Loose M, Tyson JR, de Cesare M, Brown BL, Jain M, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. F1000Research. 2015;4. doi:10.12688/f1000research.7201.1
140. Cretu Stancu M, van Roosmalen MJ, Renkens I, Nieboer MM, Middelkamp S, de Ligt J, et al. Mapping and phasing of structural variation in patient genomes using nanopore sequencing. Nat Commun. 2017;8. doi:10.1038/s41467-017-01343-4
141. David M, Dursi LJ, Yao D, Boutros PC, Simpson JT. Nanocall: an open source basecaller for Oxford Nanopore sequencing data. Bioinformatics. 2017;33: 49–55. doi:10.1093/bioinformatics/btw569
142. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat Biotechnol. 2018;36: 338–345. doi:10.1038/nbt.4060
143. Boža V, Brejová B, Vinař T. DeepNano: Deep recurrent neural networks for base calling in MinION nanopore reads. PLOS ONE. 2017;12: e0178751. doi:10.1371/journal.pone.0178751
144. Tyson JR, O'Neil NJ, Jain M, Olsen HE, Hieter P, Snutch TP. MinION-based long-read sequencing and assembly extends the *Caenorhabditis elegans* reference genome. Genome Res. 2018;28: 266–274. doi:10.1101/gr.221184.117
145. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Hum Mol Genet. 2018;27: R234–R241. doi:10.1093/hmg/ddy177
146. Tyler AD, Mataseje L, Urfano CJ, Schmidt L, Antonation KS, Mulvey MR, et al. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. Sci Rep. 2018;8: 10931. doi:10.1038/s41598-018-29334-5
147. Erlich Y. A vision for ubiquitous sequencing. Genome Res. 2015;25: 1411–1416. doi:10.1101/gr.191692.115
148. Ashton PM, Nair S, Dallman T, Rubino S, Rabsch W, Mwaigwisya S, et al. MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island. Nat Biotechnol. 2015;33: 296–300. doi:10.1038/nbt.3103
149. Faria NR, Sabino EC, Nunes MRT, Alcantara LCJ, Loman NJ, Pybus OG. Mobile real-time surveillance of Zika virus in Brazil. Genome Med. 2016;8: 97. doi:10.1186/s13073-016-0356-2
150. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature. 2016;530: 228–232. doi:10.1038/nature16996
151. Pennisi E. Pocket DNA sequencers make real-time diagnostics a reality. Science. 2016;351: 800–801. doi:10.1126/science.351.6275.800
152. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 2008;320: 1344–1349. doi:10.1126/science.1158441
153. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5: 621–628. doi:10.1038/nmeth.1226
154. Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science. 1991;251: 767–773.
155. van Hal NLW, Vorst O, van Houwelingen AMML, Kok EJ, Peijnenburg A, Aharoni A, et al. The application of DNA microarrays in gene expression analysis. J Biotechnol. 2000;78: 271–280. doi:10.1016/S0168-1656(00)00204-2
156. Cheng B, Furtado A, Henry RJ. Long-read sequencing of the coffee bean transcriptome reveals the diversity of full-length transcripts. GigaScience. 2017;6. doi:10.1093/gigascience/gix086
157. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, Nikaido I. Single-cell full-length total RNA sequencing uncovers dynamics of recursive splicing and enhancer RNAs. Nat Commun. 2018;9: 619. doi:10.1038/s41467-018-02866-0
158. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinforma Oxf Engl. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635
159. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12: 357–360. doi:10.1038/nmeth.3317
160. Schliesky S, Gowik U, Weber APM, Bräutigam A. RNA-Seq Assembly - Are We There Yet? Front Plant Sci. 2012;3: 220. doi:10.3389/fpls.2012.00220

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

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

163. Byrne SL, Nagy I, Pfeifer M, Armstead I, Swain S, Studer B, et al. A synteny-based draft genome sequence of the forage grass *Lolium perenne*. Plant J. 2015;84: 816–826. doi:10.1111/tpj.13037

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

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

166. Rio DC, Ares M, Hannon GJ, Nilsen TW. Enrichment of Poly(A)+ mRNA Using Immobilized Oligo(dT). Cold Spring Harb Protoc. 2010;2010: pdb.prot5454. doi:10.1101/pdb.prot5454

167. Krug MS, Berg SL. First-strand cDNA synthesis primed with oligo(dT). 1987; Available: https://inis.iaea.org/search/searchsinglerecord.aspx?recordsFor=SingleRecord&RN=20026117

168. Armour CD, Castle JC, Chen R, Babak T, Loerch P, Jackson S, et al. Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. Nat Methods. 2009;6: 647–649. doi:10.1038/nmeth.1360

169. Ozsolak F, Kapranov P, Foissac S, Kim SW, Fishilevich E, Monaghan AP, et al. Comprehensive Polyadenylation Site Maps in Yeast and Human Reveal Pervasive Alternative Polyadenylation. Cell. 2010;143: 1018–1029. doi:10.1016/j.cell.2010.11.020

170. O'Neil D, Glowatz H, Schlumpberger M. Ribosomal RNA depletion for efficient use of RNA-seq capacity. Curr Protoc Mol Biol. 2013;Chapter 4: Unit 4.19. doi:10.1002/0471142727.mb0419s103

171. Petrova OE, Garcia-Alcalde F, Zampaloni C, Sauer K. Comparative evaluation of rRNA depletion procedures for the improved analysis of bacterial biofilm and mixed pathogen culture transcriptomes. Sci Rep. 2017;7: 41114. doi:10.1038/srep41114

172. Hagemann-Jensen M, Abdullayev I, Sandberg R, Faridani OR. Small-seq for single-cell small-RNA sequencing. Nat Protoc. 2018;13: 2407. doi:10.1038/s41596-018-0049-y

173. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, et al. Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. Nature. 2008;453: 1239–1243. doi:10.1038/nature07002

174. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462: 315–322. doi:10.1038/nature08514

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

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

177. Sutton GG, White O, Adams MD, Kerlavage AR. TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. Genome Sci Technol. 1995;1: 9–19. doi:10.1089/gst.1995.1.9

178. Myers EW, Sutton GG, Delcher AL, Dew IM, Fasulo DP, Flanigan MJ, et al. A whole-genome assembly of *Drosophila*. Science. 2000;287: 2196–2204.

179. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291: 1304–1351. doi:10.1126/science.1058040

180. Huang X, Madan A. CAP3: A DNA sequence assembly program. Genome Res. 1999;9: 868–877.

181. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, et al. ARACHNE: a whole-genome shotgun assembler. Genome Res. 2002;12: 177–189. doi:10.1101/gr.208902

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

183. Pevzner PA. 1-Tuple DNA Sequencing: Computer Analysis. J Biomol Struct Dyn. 1989;7: 63–73. doi:10.1080/07391102.1989.10507752

184. Idury RM, Waterman MS. A new algorithm for DNA sequence assembly. J Comput Biol J Comput Mol Cell Biol. 1995;2: 291–306. doi:10.1089/cmb.1995.2.291

185. Pevzner PA, Tang H, Waterman MS. An Eulerian path approach to DNA fragment assembly. Proc Natl Acad Sci. 2001;98: 9748–9753. doi:10.1073/pnas.171285098

186. Myers EW. Toward Simplifying and Accurately Formulating Fragment Assembly. J Comput Biol. 1995;2: 275–290. doi:10.1089/cmb.1995.2.275

187. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5: R12. doi:10.1186/gb-2004-5-2-r12

188. Paszkiewicz K, Studholme DJ. *De novo* assembly of short sequence reads. Brief Bioinform. 2010;11: 457–472. doi:10.1093/bib/bbq020

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

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

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

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

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

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

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

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

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

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

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

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

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

202. Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat Methods. 2013;10: 563–569. doi:10.1038/nmeth.2474

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

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

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

206. Krishnakumar R, Sinha A, Bird SW, Jayamohan H, Edwards HS, Schoeniger JS, et al. Systematic and stochastic influences on the performance of the MinION nanopore sequencer across a range of nucleotide bias. Sci Rep. 2018;8: 3159. doi:10.1038/s41598-018-21484-w

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

208. Salmela L, Walve R, Rivals E, Ukkonen E. Accurate self-correction of errors in long reads using de Bruijn graphs. Bioinformatics. 2017;33: 799–806. doi:10.1093/bioinformatics/btw321

209. Bao E, Lan L. HALC: High throughput algorithm for long read error correction. BMC Bioinformatics. 2017;18: 204. doi:10.1186/s12859-017-1610-3

210. Choudhury O, Chakrabarty A, Emrich SJ. HECIL: A Hybrid Error Correction Algorithm for Long Reads with Iterative Learning. Sci Rep. 2018;8: 9936. doi:10.1038/s41598-018-28364-3

211. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33: 623–630. doi:10.1038/nbt.3238

212. Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, et al. Chromosome-level assembly of *Arabidopsis thaliana* L*er* reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci. 2016;113: E4052–E4060. doi:10.1073/pnas.1607532113

213. Pucker B, Holtgraewe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, et al. A Chromosome-level Sequence Assembly Reveals the Structure of the *Arabidopsis thaliana* Nd-1 Genome and its Gene Set. bioRxiv. 2018; doi:https://doi.org/10.1101/407627

214. Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, et al. *De novo* assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore MinION sequencer. GigaScience. 2017;6: 1–13. doi:10.1093/gigascience/giw018

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

216. Kolmogorov M, Yuan J, Lin Y, Pevzner P. Assembly of Long Error-Prone Reads Using Repeat Graphs. bioRxiv. 2018; 247148. doi:10.1101/247148

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

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

219. Saint-Oyant LH, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, et al. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. Nat Plants. 2018;4: 473. doi:10.1038/s41477-018-0166-1

220. Watson M. Mind the gaps - ignoring errors in long read assemblies critically affects protein prediction. bioRxiv. 2018; 285049. doi:10.1101/285049

221. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463: 178–183. doi:10.1038/nature08670

222. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet. 2011;43: 109–116. doi:10.1038/ng.740

223. Nagarajan N, Pop M. Sequence assembly demystified. Nat Rev Genet. 2013;14: 157–167. doi:10.1038/nrg3367

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

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

226. Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ, Wang YK. Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. Science. 1993;262: 110–114. doi:10.1126/science.8211116

227. Lam ET, Hastie A, Lin C, Ehrlich D, Das SK, Austin MD, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. Nat Biotechnol. 2012;30: 771–776. doi:10.1038/nbt.2303

228. Valouev A, Zhang Y, Schwartz DC, Waterman MS. Refinement of optical map assemblies. Bioinforma Oxf Engl. 2006;22: 1217–1224. doi:10.1093/bioinformatics/btl063

229. Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nat Genet. 2016;48: 1225–1232. doi:10.1038/ng.3657

230. Jarvis DE, Ho YS, Lightfoot DJ, Schmöckel SM, Li B, Borm TJA, et al. The genome of *Chenopodium quinoa*. Nature. 2017;542: 307–312. doi:10.1038/nature21370

231. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet. 1980;32: 314–331.

232. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, et al. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 1995;23: 4407–4414.

233. Maughan PJ, Saghai Maroof MA, Buss GR, Huestis GM. Amplified fragment length polymorphism (AFLP) in soybean: species diversity, inheritance, and near-isogenic line analysis. TAG Theor Appl Genet Theor Angew Genet. 1996;93: 392–401. doi:10.1007/BF00223181

234. Savelkoul PHM, Aarts HJM, de Haas J, Dijkshoorn L, Duim B, Otsen M, et al. Amplified-Fragment Length Polymorphism Analysis: the State of an Art. J Clin Microbiol. 1999;37: 3083–3091.

235. Bonin A, Ehrich D, Manel S. Statistical analysis of amplified fragment length polymorphism data: a toolbox for molecular ecologists and evolutionists. Mol Ecol. 2007;16: 3737–3758. doi:10.1111/j.1365-294X.2007.03435.x

236. Zietkiewicz E, Rafalski A, Labuda D. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics. 1994;20: 176–183. doi:10.1006/geno.1994.1151

237. Faris JD, Haen KM, Gill BS. Saturation mapping of a gene-rich recombination hot spot region in wheat. Genetics. 2000;154: 823–835.

238. Yao H, Zhou Q, Li J, Smith H, Yandeau M, Nikolau BJ, et al. Molecular characterization of meiotic recombination across the 140-kb multigenic a1-sh2 interval of maize. Proc Natl Acad Sci. 2002;99: 6157–6162. doi:10.1073/pnas.082562199

239. Ko Y-J, Kim JS, Kim S. misMM: An Integrated Pipeline for Misassembly Detection Using Genotyping-by-Sequencing and Its Validation with BAC End Library Sequences and Gene Synteny. Genomics Inform. 2017;15: 128–135. doi:10.5808/GI.2017.15.4.128

240. Meader S, Hillier LW, Locke D, Ponting CP, Lunter G. Genome assembly quality: assessment and improvement using the neutral indel model. Genome Res. 2010;20: 675–684. doi:10.1101/gr.096966.109

241. Zhang W, Chen J, Yang Y, Tang Y, Shang J, Shen B. A Practical Comparison of *De Novo* Genome Assembly Software Tools for Next-Generation Sequencing Technologies. PLoS ONE. 2011;6. doi:10.1371/journal.pone.0017915

242. Earl D, Bradnam K, St John J, Darling A, Lin D, Fass J, et al. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. Genome Res. 2011;21: 2224–2241. doi:10.1101/gr.126599.111

243. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409: 860–921. doi:10.1038/35057062

244. Salzberg SL, Yorke JA. Beware of mis-assembled genomes. Bioinformatics. 2005;21: 4320–4321. doi:10.1093/bioinformatics/bti769

245. Milne I, Bayer M, Cardle L, Shaw P, Stephen G, Wright F, et al. Tablet—next generation sequence assembly visualization. Bioinformatics. 2010;26: 401–402. doi:10.1093/bioinformatics/btp666

246. Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol. 2014;18: 31–36. doi:10.1016/j.pbi.2014.01.003

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

248. Li R, Fan W, Tian G, Zhu H, He L, Cai J, et al. The sequence and *de novo* assembly of the giant panda genome. Nature. 2010;463: 311–317. doi:10.1038/nature08696

249. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998;8: 186–194.

250. Gritsenko AA, Nijkamp JF, Reinders MJT, de Ridder D. GRASS: a generic algorithm for scaffolding next-generation sequencing assemblies. Bioinforma Oxf Engl. 2012;28: 1429–1437. doi:10.1093/bioinformatics/bts175

251. Hunt M, Newbold C, Berriman M, Otto TD. A comprehensive evaluation of assembly scaffolding tools. Genome Biol. 2014;15: R42. doi:10.1186/gb-2014-15-3-r42

252. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2012;13: 146. doi:10.1038/nrg3164

253. Church DM, Goodstadt L, Hillier LW, Zody MC, Goldstein S, She X, et al. Lineage-Specific Biology Revealed by a Finished Genome Assembly of the Mouse. PLOS Biol. 2009;7: e1000112. doi:10.1371/journal.pbio.1000112

254. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, et al. Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc Natl Acad Sci U S A. 2011;108: 10249–10254. doi:10.1073/pnas.1107739108

255. Arumuganathan K, Earle ED. Nuclear DNA content of some important plant species. Plant Mol Biol Report. 1991;9: 208–218. doi:10.1007/BF02672069

256. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27: 764–770. doi:10.1093/bioinformatics/btr011

257. Vurture GW, Sedlazeck FJ, Nattestad M, Underwood CJ, Fang H, Gurtowski J, et al. GenomeScope: fast reference-free genome profiling from short reads. Bioinformatics. 2017;33: 2202–2204. doi:10.1093/bioinformatics/btx153

258. Sun H, Ding J, Piednoël M, Schneeberger K. findGSE: estimating genome size variation within human and *Arabidopsis* using k-mer frequencies. Bioinforma Oxf Engl. 2018;34: 550–557. doi:10.1093/bioinformatics/btx637

259. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29: 1072–1075. doi:10.1093/bioinformatics/btt086

260. Khelik K, Lagesen K, Sandve GK, Rognes T, Nederbragt AJ. NucDiff: in-depth characterization and annotation of differences between two sets of DNA sequences. BMC Bioinformatics. 2017;18: 338. doi:10.1186/s12859-017-1748-z

261. Delcher AL, Phillippy A, Carlton J, Salzberg SL. Fast algorithms for large-scale genome alignment and comparison. Nucleic Acids Res. 2002;30: 2478–2483.

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

263. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic Acids Res. 2017;45: D744–D749. doi:10.1093/nar/gkw1119

264. Kurtz S, Narechania A, Stein JC, Ware D. A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 2008;9: 517. doi:10.1186/1471-2164-9-517

265. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18: 188–196. doi:10.1101/gr.6743907

266. Furnham N, de Beer TAP, Thornton JM. Current challenges in genome annotation through structural biology and bioinformatics. Curr Opin Struct Biol. 2012;22: 594–601. doi:10.1016/j.sbi.2012.07.005

267. König S, Romoth L, Stanke M. Comparative Genome Annotation. In: Setubal JC, Stoye J, Stadler PF, editors. Comparative Genomics: Methods and Protocols. New York, NY: Springer New York; 2018. pp. 189–212. doi:10.1007/978-1-4939-7463-4_6

268. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinforma Oxf Engl. 2003;19 Suppl 2: ii215-225.

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

270. Cook D, Valle-Inclan JE, Pajoro A, Rovenich H, Thomma B, Faino L. Long Read Annotation (LoReAn): automated eukaryotic genome annotation based on long-read cDNA sequencing. Plant Physiol. 2019; pp.00848.2018. doi:10.1104/pp.18.00848

271. Stormo GD. Gene-Finding Approaches for Eukaryotes. Genome Res. 2000;10: 394–397. doi:10.1101/gr.10.4.394

272. Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005;33: W465–W467. doi:10.1093/nar/gki458

273. Stormo GD, Haussler D. Optimally parsing a sequence into different classes based on multiple types of evidence. Proc Int Conf Intell Syst Mol Biol. 1994;2: 369–375.

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

275. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinforma Oxf Engl. 2011;27: 757–763. doi:10.1093/bioinformatics/btr010

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

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

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

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

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

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

282. Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E. The Ensembl Core Software Libraries. Genome Res. 2004;14: 929–933. doi:10.1101/gr.1857204

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

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

285. Coward E, Haas SA, Vingron M. SpliceNest: visualizing gene structure and alternative splicing based on EST clusters. Trends Genet. 2002;18: 53–55. doi:10.1016/S0168-9525(01)02525-2

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

287. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

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

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

290. Keibler E, Brent MR. Eval: A software package for analysis of genome annotations. BMC Bioinformatics. 2003;4: 50. doi:10.1186/1471-2105-4-50

291. Allen JE, Pertea M, Salzberg SL. Computational Gene Prediction Using Multiple Sources of Evidence. Genome Res. 2004;14: 142–148. doi:10.1101/gr.1562804

292. Standage DS, Brendel VP. ParsEval: parallel comparison and analysis of gene structure annotations. BMC Bioinformatics. 2012;13: 187. doi:10.1186/1471-2105-13-187

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

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

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

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

297. Britten RJ, Kohne DE. Repeated Sequences in DNA. Science. 1968;161: 529–540. doi:10.1126/science.161.3841.529

298. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. Nat Genet. 1998;20: 43–45. doi:10.1038/1695

299. Vitte C, Fustier M-A, Alix K, Tenaillon MI. The bright side of transposons in crop evolution. Brief Funct Genomics. 2014;13: 276–295. doi:10.1093/bfgp/elu002

300. Volff J-N. Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. BioEssays News Rev Mol Cell Dev Biol. 2006;28: 913–922. doi:10.1002/bies.20452

301. Martin A, Troadec C, Boualem A, Rajab M, Fernandez R, Morin H, et al. A transposon-induced epigenetic change leads to sex determination in melon. Nature. 2009;461: 1135–1138. doi:10.1038/nature08498

302. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nat Genet. 2011;43: 1160–1163. doi:10.1038/ng.942

303. Butelli E, Licciardello C, Zhang Y, Liu J, Mackay S, Bailey P, et al. Retrotransposons Control Fruit-Specific, Cold-Dependent Accumulation of Anthocyanins in Blood Oranges. Plant Cell. 2012;24: 1242–1255. doi:10.1105/tpc.111.095232

304. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13: R107. doi:10.1186/gb-2012-13-11-r107

305. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable Elements Are Major Contributors to the Origin, Diversification, and Regulation of Vertebrate Long Noncoding RNAs. PLoS Genet. 2013;9. doi:10.1371/journal.pgen.1003470

306. Hoen DR, Bureau TE. Discovery of novel genes derived from transposable elements using integrative genomic analysis. Mol Biol Evol. 2015;32: 1487–1506. doi:10.1093/molbev/msv042

307. Jouffroy O, Saha S, Mueller L, Quesneville H, Maumus F. Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. BMC Genomics. 2016;17. doi:10.1186/s12864-016-2980-z

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

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

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

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

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

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

314. Nikonorova N, Yue K, Beeckman T, De Smet I. *Arabidopsis* research requires a critical re-evaluation of genetic tools. J Exp Bot. 2018;69: 3541–3544. doi:10.1093/jxb/ery161

315. Nordborg M, Borevitz JO, Bergelson J, Berry CC, Chory J, Hagenblad J, et al. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet. 2002;30: 190–193. doi:10.1038/ng813

316. Nordborg M, Hu TT, Ishino Y, Jhaveri J, Toomajian C, Zheng H, et al. The Pattern of Polymorphism in *Arabidopsis thaliana*. PLOS Biol. 2005;3: e196. doi:10.1371/journal.pbio.0030196

317. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common Sequence Polymorphisms Shaping Genetic Diversity in *Arabidopsis thaliana*. Science. 2007;317: 338–342. doi:10.1126/science.1138632

318. Kim S, Plagnol V, Hu TT, Toomajian C, Clark RM, Ossowski S, et al. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. Nat Genet. 2007;39: 1151–1155. doi:10.1038/ng2115

319. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res. 2008;18: 2024–2033. doi:10.1101/gr.080200.108

320. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet. 2011;43: 956–963. doi:10.1038/ng.911

321. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature. 2011;477: 419–423. doi:10.1038/nature10414

322. Horton MW, Hancock AM, Huang YS, Toomajian C, Atwell S, Auton A, et al. Genome-wide patterns of genetic variation in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. Nat Genet. 2012;44: 212–216. doi:10.1038/ng.1042

323. Schmitz RJ, Schultz MD, Urich MA, Nery JR, Pelizzola M, Libiger O, et al. Patterns of Population Epigenomic Diversity. Nature. 2013;495: 193–198. doi:10.1038/nature11968

324. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive genomic variation and strong selection in *Arabidopsis thaliana* lines from Sweden. Nat Genet. 2013;45: 884–890. doi:10.1038/ng.2678

325. Hagmann J, Becker C, Müller J, Stegle O, Meyer RC, Wang G, et al. Century-scale Methylome Stability in a Recently Diverged *Arabidopsis thaliana* Lineage. PLOS Genet. 2015;11: e1004920. doi:10.1371/journal.pgen.1004920

326. Zou Y-P, Hou X-H, Wu Q, Chen J-F, Li Z-W, Han T-S, et al. Adaptation of *Arabidopsis thaliana* to the Yangtze River basin. Genome Biol. 2017;18. doi:10.1186/s13059-017-1378-9

327. Durvasula A, Fulgione A, Gutaker RM, Alacakaptan SI, Flood PJ, Neto C, et al. African genomes illuminate the early history and transition to selfing in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A. 2017;114: 5213–5218. doi:10.1073/pnas.1616736114

328. Schmid KJ, Sörensen TR, Stracke R, Törjék O, Altmann T, Mitchell-Olds T, et al. Large-Scale Identification and Analysis of Genome-Wide Single-Nucleotide Polymorphisms for Mapping in *Arabidopsis thaliana*. Genome Res. 2003;13: 1250–1257. doi:10.1101/gr.728603

329. Platt A, Horton M, Huang YS, Li Y, Anastasio AE, Mulyati NW, et al. The Scale of Population Structure in *Arabidopsis thaliana*. PLoS Genet. 2010;6. doi:10.1371/journal.pgen.1000843

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

331. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012;490: 497–501. doi:10.1038/nature11532

332. Aflitos S, Schijlen E, Jong H de, Ridder D de, Smit S, Finkers R, et al. Exploring genetic variation in the tomato (*Solanum* section Lycopersicon) clade by whole-genome sequencing. Plant J. 2014;80: 136–148. doi:10.1111/tpj.12616

333. Hazzouri KM, Flowers JM, Visser HJ, Khierallah HSM, Rosas U, Pham GM, et al. Whole genome re-sequencing of date palms yields insights into diversification of a fruit tree crop. Nat Commun. 2015;6: 8824. doi:10.1038/ncomms9824

334. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33: 408–414. doi:10.1038/nbt.3096

335. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557: 43. doi:10.1038/s41586-018-0063-9

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

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

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

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

340. Zhao X, Weber AM, Mills RE. A recurrence-based approach for validating structural variation using long-read sequencing technology. GigaScience. 2017;6: 1–9. doi:10.1093/gigascience/gix061

341. Sedlazeck FJ, Lee H, Darby CA, Schatz MC. Piercing the dark matter: bioinformatics of long-range sequencing and mapping. Nat Rev Genet. 2018;19: 329. doi:10.1038/s41576-018-0003-4

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

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

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

345. Laibach F. *Arabidopsis thaliana* (L.) Heynh. als Objekt für genetische und entwicklungsphysiologische Untersuchungen. Bot Arch. 1943;44: 439–455.

346. Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M. *Arabidopsis thaliana*: a model plant for genome analysis. Science. 1998;282: 662, 679–682.

347. Meyerowitz EM. Prehistory and History of Arabidopsis Research. Plant Physiol. 2001;125: 15–19. doi:10.1104/pp.125.1.15

348. Koornneef M, Meinke D. The development of *Arabidopsis* as a model plant. Plant J. 2010;61: 909–921. doi:10.1111/j.1365-313X.2009.04086.x

349. Berardini TZ, Reiser L, Li D, Mezheritsky Y, Muller R, Strait E, et al. The *Arabidopsis* Information Resource: Making and Mining the 'Gold Standard' Annotated Reference Plant Genome. Genes N Y N 2000. 2015;53: 474–485. doi:10.1002/dvg.22877

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

351. Van Norman JM, Benfey PN. *Arabidopsis thaliana* as a Model Organism in Systems Biology. Wiley Interdiscip Rev Syst Biol Med. 2009;1: 372–379. doi:10.1002/wsbm.25

352. Laibach F. Zur Frage nach der Individualität der Chromosomen im Pflanzenreich. 1907; Available: https://www.biodiversitylibrary.org/item/27073#page/233/mode/1up

353. Laibach F. 60 Jahre *Arabidopsis*-Forschung. 1965; Available: https://www.arabidopsis.org/ais/1965/laiba-1965-aagle.html

354. Clough SJ, Bent AF. Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*. Plant J Cell Mol Biol. 1998;16: 735–743.

355. Koch MA. The plant model system *Arabidopsis* set in an evolutionary, systematic, and spatio-temporal context. J Exp Bot. 2019;70: 55–67. doi:10.1093/jxb/ery340

356. Darwin C. The Effects of Cross and Self Fertilisation in the Vegetable Kingdom by Charles Darwin [Internet]. Cambridge; 1876. doi:10.1017/CBO9780511694202

357. Stebbins GL. Variation and evolution in plants. Geoffrey Cumberlege.; London; 1950.

358. Takebayashi N, Morrell PL. Is Self-Fertilization an Evolutionary Dead End? Revisiting an Old Hypothesis with Genetic Theories and a Macroevolutionary Approach. Am J Bot. 2001;88: 1143–1150. doi:10.2307/3558325

359. Igic B, Busch JW. Is self-fertilization an evolutionary dead end? New Phytol. 2013;198: 386–397. doi:10.1111/nph.12182

360. Wright SI, Kalisz S, Slotte T. Evolutionary consequences of self-fertilization in plants. Proc R Soc B Biol Sci. 2013;280. doi:10.1098/rspb.2013.0133

361. Veiga RSL, Faccio A, Genre A, Pieterse CMJ, Bonfante P, Heijden MGA van der. Arbuscular mycorrhizal fungi reduce growth and infect roots of the non-host plant *Arabidopsis thaliana*. Plant Cell Environ. 2013;36: 1926–1937. doi:10.1111/pce.12102

362. Hiruma K, Gerlach N, Sacristán S, Nakano RT, Hacquard S, Kracher B, et al. Root Endophyte *Colletotrichum tofieldiae* Confers Plant Fitness Benefits that Are Phosphate Status Dependent. Cell. 2016;165: 464–474. doi:10.1016/j.cell.2016.02.028

363. Behnke N, Suprianto E, Möllers C. A major QTL on chromosome C05 significantly reduces acid detergent lignin (ADL) content and increases seed oil and protein content in oilseed rape (*Brassica napus* L.). Theor Appl Genet. 2018;131: 2477–2492. doi:10.1007/s00122-018-3167-6

364. phyloT: a phylogenetic tree generator [Internet]. 2018 [cited 5 Jan 2019]. Available: https://phylot.biobyte.de/

365. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res. 2016;44: W242–W245. doi:10.1093/nar/gkw290

366. Rédei GP. Supervital Mutants of *Arabidopsis*. Genetics. 1962;47: 443–460.

367. Rédei GP. Single locus heterosis. Z Für Vererbungslehre. 1962;93: 164–170. doi:10.1007/BF00897025

368. Rédei GP. A heuristic glance at the past of *Arabidopsis* genetics. Methods in Arabidopsis Research. WORLD SCIENTIFIC; 1992. pp. 1–15. doi:10.1142/9789814439701_0001

369. Kranz AR, Kirchheim B. Genetic resources in *Arabidopsis*. Arab Inf Serv. 1987; Available: http://agris.fao.org/agris-search/search.do?recordID=US201301398043

370. Deslandes L, Pileur F, Liaubet L, Camut S, Can C, Williams K, et al. Genetic Characterization of *RRS1*, a Recessive Locus in *Arabidopsis thaliana* that Confers Resistance to the Bacterial Soilborne Pathogen *Ralstonia solanacearum*. Mol Plant Microbe Interact. 1998;11: 659–667. doi:10.1094/MPMI.1998.11.7.659

371. Yang C-H, Ho G-D. Resistance and susceptibility of *Arabidopsis thaliana* to bacterial wilt caused by *Ralstonia solanacearum*. Phytopathology. 1998;88: 330–334.

372. Meyer RC, Törjék O, Becher M, Altmann T. Heterosis of Biomass Production in *Arabidopsis*. Establishment during Early Development. Plant Physiol. 2004;134: 1813–1823. doi:10.1104/pp.103.033001

373. Werner JD, Borevitz JO, Warthmann N, Trainer GT, Ecker JR, Chory J, et al. Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. Proc Natl Acad Sci. 2005;102: 2460–2465. doi:10.1073/pnas.0409474102

374. Gonzalez A, Zhao M, Leavitt JM, Lloyd AM. Regulation of the anthocyanin biosynthetic pathway by the TTG1/bHLH/Myb transcriptional complex in *Arabidopsis* seedlings. Plant J. 2008;53: 814–827. doi:10.1111/j.1365-313X.2007.03373.x

375. Ishihara H, Tohge T, Viehöver P, Fernie AR, Weisshaar B, Stracke R. Natural variation in flavonol accumulation in *Arabidopsis* is determined by the flavonol glucosyltransferase BGLU6. J Exp Bot. 2016;67: 1505–1517. doi:10.1093/jxb/erv546

376. 1001 Genomes [Internet]. [cited 3 Jan 2019]. Available: http://1001genomes.org/

377. Chang C, Bowman JL, DeJohn AW, Lander ES, Meyerowitz EM. Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. Proc Natl Acad Sci. 1988;85: 6856–6860. doi:10.1073/pnas.85.18.6856

378. Weigel D. Natural Variation in *Arabidopsis*: From Molecular Genetics to Ecological Genomics. Plant Physiol. 2012;158: 2–22. doi:10.1104/pp.111.189845

379. Falconer DS. Introduction to Quantitative Genetics. Longman; 1960.

380. Lister C, Dean C. Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana*. Plant J. 1993;4: 745–750. doi:10.1046/j.1365-313X.1993.04040745.x

381. Koornneef M, Alonso-Blanco C, Vreugdenhil D. Naturally occurring genetic variation in *Arabidopsis thaliana*. Annu Rev Plant Biol. 2004;55: 141–172. doi:10.1146/annurev.arplant.55.031903.141605

382. Bergelson J, Roux F. Towards identifying genes underlying ecologically relevant traits in *Arabidopsis thaliana*. Nat Rev Genet. 2010;11: 867–879. doi:10.1038/nrg2896

383. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, et al. Association Mapping: Critical Considerations Shift from Genotyping to Experimental Design. Plant Cell. 2009;21: 2194–2202. doi:10.1105/tpc.109.068437

384. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40: D1202–D1210. doi:10.1093/nar/gkr1090

385. Emrich SJ, Barbazuk WB, Li L, Schnable PS. Gene discovery and annotation using LCM-454 transcriptome sequencing. Genome Res. 2007;17: 69–73. doi:10.1101/gr.5145806

386. Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. Sampling the *Arabidopsis* Transcriptome with Massively Parallel Pyrosequencing. Plant Physiol. 2007;144: 32–42. doi:10.1104/pp.107.096677

387. Kawakatsu T, Huang SC, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. Cell. 2016;166: 492–505. doi:10.1016/j.cell.2016.06.044

388. Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, et al. Dissecting a hidden gene duplication: the *Arabidopsis thaliana SEC10* locus. PloS One. 2014;9: e94077. doi:10.1371/journal.pone.0094077

389. Noël L, Moores TL, Biezen EA van der, Parniske M, Daniels MJ, Parker JE, et al. Pronounced Intraspecific Haplotype Divergence at the *RPP5* Complex Disease Resistance Locus of *Arabidopsis*. Plant Cell. 1999;11: 2099–2111. doi:10.1105/tpc.11.11.2099

390. Belser C, Istace B, Denis E, Dubarry M, Baurens F-C, Falentin C, et al. Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. Nat Plants. 2018;4: 879. doi:10.1038/s41477-018-0289-4

391. Kagale S, Robinson SJ, Nixon J, Xiao R, Huebert T, Condie J, et al. Polyploid Evolution of the Brassicaceae during the Cenozoic Era. Plant Cell. 2014;26: 2777–2791. doi:10.1105/tpc.114.126391

392. Murat F, Louis A, Maumus F, Armero A, Cooke R, Quesneville H, et al. Understanding Brassicaceae evolution through ancestral genome reconstruction. Genome Biol. 2015;16: 262. doi:10.1186/s13059-015-0814-y

393. Buisine N, Quesneville H, Colot V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. Genomics. 2008;91: 467–475. doi:10.1016/j.ygeno.2008.01.005

394. Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci U S A. 2011;108: 2322–2327. doi:10.1073/pnas.1018222108

395. de la Chaux N, Tsuchimatsu T, Shimizu KK, Wagner A. The predominantly selfing plant *Arabidopsis thaliana* experienced a recent reduction in transposable element abundance compared to its outcrossing relative *Arabidopsis lyrata*. Mob DNA. 2012;3: 2. doi:10.1186/1759-8753-3-2

396. Maere S, De Bodt S, Raes J, Casneuf T, Van Montagu M, Kuiper M, et al. Modeling gene and genome duplications in eukaryotes. Proc Natl Acad Sci U S A. 2005;102: 5454–5459. doi:10.1073/pnas.0501102102

397. Wolfe K. Robustness—it's not where you think it is. Nat Genet. 2000;25: 3–4. doi:10.1038/75560

398. Bowers JE, Chapman BA, Rong J, Paterson AH. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature. 2003;422: 433–438. doi:10.1038/nature01521

399. Charlesworth D, Vekemans X. How and when did *Arabidopsis thaliana* become highly self-fertilising. BioEssays News Rev Mol Cell Dev Biol. 2005;27: 472–476. doi:10.1002/bies.20231

400. Nordborg M. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial self-fertilization. Genetics. 2000;154: 923–929.

401. Koch MA, Haubold B, Mitchell-Olds T. Comparative Evolutionary Analysis of Chalcone Synthase and Alcohol Dehydrogenase Loci in *Arabidopsis*, *Arabis*, and Related Genera (Brassicaceae). Mol Biol Evol. 2000;17: 1483–1498. doi:10.1093/oxfordjournals.molbev.a026248

402. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The Rate and Molecular Spectrum of Spontaneous Mutations in *Arabidopsis thaliana*. Science. 2010;327. doi:10.1126/science.1180677

403. Exposito-Alonso M, Becker C, Schuenemann VJ, Reiter E, Setzer C, Slovak R, et al. The rate and potential relevance of new mutations in a colonizing plant lineage. bioRxiv. 2017; 050203. doi:10.1101/050203

404. Yang Y-F, Zhu T, Niu D-K. Association of Intron Loss with High Mutation Rate in *Arabidopsis*: Implications for Genome Size Evolution. Genome Biol Evol. 2013;5: 723–733. doi:10.1093/gbe/evt043

405. Payne BL, Alvarez-Ponce D. Higher Rates of Protein Evolution in the Self-Fertilizing Plant *Arabidopsis thaliana* than in the Out-Crossers *Arabidopsis lyrata* and *Arabidopsis halleri*. Genome Biol Evol. 2018;10: 895–900. doi:10.1093/gbe/evy053

406. Pollak E. On the Theory of Partially Inbreeding Finite Populations. I. Partial Selfing. Genetics. 1987;117: 353–360.

407. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nat Rev Genet. 2009;10: 195–205. doi:10.1038/nrg2526

408. Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983.

409. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. The dynamic ups and downs of genome size evolution in Brassicaceae. Mol Biol Evol. 2009;26: 85–98. doi:10.1093/molbev/msn223

410. Schranz ME, Lysak MA, Mitchell-Olds T. The ABC's of comparative genomics in the Brassicaceae: building blocks of crucifer genomes. Trends Plant Sci. 2006;11: 535–542. doi:10.1016/j.tplants.2006.09.002

411. Koenig D, Weigel D. Beyond the thale: comparative genomics and genetics of *Arabidopsis* relatives. Nat Rev Genet. 2015;16: 285–298. doi:10.1038/nrg3883

412. Ahmed I, Sarazin A, Bowler C, Colot V, Quesneville H. Genome-wide evidence for local DNA methylation spreading from small RNA-targeted sequences in *Arabidopsis*. Nucleic Acids Res. 2011;39: 6919–6931. doi:10.1093/nar/gkr324

413. Slotte T, Hazzouri KM, Ågren JA, Koenig D, Maumus F, Guo Y-L, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet. 2013;45: 831–835. doi:10.1038/ng.2669

414. Schmuths H, Meister A, Horres R, Bachmann K. Genome Size Variation among Accessions of *Arabidopsis thaliana*. Ann Bot. 2004;93: 317–321. doi:10.1093/aob/mch037

415. Davison J, Tyagi A, Comai L. Large-scale polymorphism of heterochromatic repeats in the DNA of *Arabidopsis thaliana*. BMC Plant Biol. 2007;7: 44. doi:10.1186/1471-2229-7-44

416. Leutwiler LS, Hough-Evans BR, Meyerowitz EM. The DNA of *Arabidopsis thaliana*. Mol Gen Genet MGG. 1984;194: 15–23. doi:10.1007/BF00383491

417. Francis DM, Hulbert SH, Michelmore RW. Genome size and complexity of the obligate fungal pathogen, *Bremia lactucae*. Exp Mycol. 1990;14: 299–309.

418. Bennett MD, Smith JB. Nuclear DNA Amounts in Angiosperms. Philos Trans Biol Sci. 1991;334: 309–345.

419. Bennett MD, Leitch IJ. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. Ann Bot. 2011;107: 467–590. doi:10.1093/aob/mcq258

420. Bennetzen JL. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol. 2000;42: 251–269.

421. Casacuberta JM, Jackson S, Panaud O, Purugganan M, Wendel J. Evolution of Plant Phenotypes, from Genomes to Traits. G3 Genes Genomes Genet. 2016;6: 775–778. doi:10.1534/g3.115.025502

422. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng J-F, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43: 476–481. doi:10.1038/ng.807

423. Bestor TH. Sex brings transposons and genomes into conflict. In: McDonald JF, editor. Transposable Elements and Genome Evolution. Dordrecht: Springer Netherlands; 2000. pp. 289–295. doi:10.1007/978-94-011-4156-7_28

424. Meinke D, Muralla R, Sweeney C, Dickerman A. Identifying essential genes in *Arabidopsis thaliana*. Trends Plant Sci. 2008;13: 483–491. doi:10.1016/j.tplants.2008.06.003

425. Meinke DW, Meinke LK, Showalter TC, Schissel AM, Mueller LA, Tzafrir I. A Sequence-Based Map of *Arabidopsis* Genes with Mutant Phenotypes. Plant Physiol. 2003;131: 409–418. doi:10.1104/pp.014134

426. Pagnussat GC, Yu H-J, Ngo QA, Rajani S, Mayalagu S, Johnson CS, et al. Genetic and molecular identification of genes required for female gametophyte development and function in *Arabidopsis*. Development. 2005;132: 603–614. doi:10.1242/dev.01595

427. Tan S, Zhong Y, Hou H, Yang S, Tian D. Variation of presence/absence genes among *Arabidopsis* populations. BMC Evol Biol. 2012;12: 86. doi:10.1186/1471-2148-12-86

428. Contreras-Moreira B, Cantalapiedra CP, García-Pereira MJ, Gordon SP, Vogel JP, Igartua E, et al. Analysis of Plant Pan-Genomes and Transcriptomes with GET_HOMOLOGUES-EST, a Clustering Solution for Sequences of the Same Species. Front Plant Sci. 2017;8. doi:10.3389/fpls.2017.00184

429. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, et al. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome." Proc Natl Acad Sci U S A. 2005;102: 13950–13955. doi:10.1073/pnas.0506758102

430. Golicz AA, Batley J, Edwards D. Towards plant pangenomics. Plant Biotechnol J. 2016;14: 1099–1105. doi:10.1111/pbi.12499

431. Swanson-Wagner RA, Eichten SR, Kumari S, Tiffin P, Stein JC, Ware D, et al. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. Genome Res. 2010;20: 1689–1699. doi:10.1101/gr.109165.110

432. Remans T, Smeets K, Opdenakker K, Mathijsen D, Vangronsveld J, Cuypers A. Normalisation of real-time RT-PCR gene expression measurements in *Arabidopsis thaliana* exposed to increased metal concentrations. Planta. 2008;227: 1343–1349. doi:10.1007/s00425-008-0706-4

433. Udvardi MK, Czechowski T, Scheible W-R. Eleven Golden Rules of Quantitative RT-PCR. Plant Cell. 2008;20: 1736–1737. doi:10.1105/tpc.108.061143

434. Hong SM, Bahn SC, Lyu A, Jung HS, Ahn JH. Identification and Testing of Superior Reference Genes for a Starting Pool of Transcript Normalization in *Arabidopsis*. Plant Cell Physiol. 2010;51: 1694–1706. doi:10.1093/pcp/pcq128

435. Dekkers BJW, Willems L, Bassel GW, van Bolderen-Veldkamp RPM, Ligterink W, Hilhorst HWM, et al. Identification of reference genes for RT-qPCR expression analysis in *Arabidopsis* and tomato seeds. Plant Cell Physiol. 2012;53: 28–37. doi:10.1093/pcp/pcr113

436. Wisman E, Ohlrogge J. *Arabidopsis* Microarray Service Facilities. Plant Physiol. 2000;124: 1468–1471. doi:10.1104/pp.124.4.1468

437. Girke T, Todd J, Ruuska S, White J, Benning C, Ohlrogge J. Microarray Analysis of Developing *Arabidopsis* Seeds. Plant Physiol. 2000;124: 1570–1581.

438. Schenk PM, Kazan K, Wilson I, Anderson JP, Richmond T, Somerville SC, et al. Coordinated plant defense responses in *Arabidopsis* revealed by microarray analysis. Proc Natl Acad Sci U S A. 2000;97: 11655–11660.

439. Postnikova OA, Nemchinov LG. Comparative analysis of microarray data in *Arabidopsis* transcriptome during compatible interactions with plant viruses. Virol J. 2012;9: 101. doi:10.1186/1743-422X-9-101

440. Birnbaum K, Shasha DE, Wang JY, Jung JW, Lambert GM, Galbraith DW, et al. A gene expression map of the *Arabidopsis* root. Science. 2003;302: 1956–1960. doi:10.1126/science.1090022

441. Zimmermann P, Hirsch-Hoffmann M, Hennig L, Gruissem W. GENEVESTIGATOR. *Arabidopsis* Microarray Database and Analysis Toolbox. Plant Physiol. 2004;136: 2621–2632. doi:10.1104/pp.104.046367

442. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, Vingron M, et al. A gene expression map of *Arabidopsis thaliana* development. Nat Genet. 2005;37: 501–506. doi:10.1038/ng1543

443. Nakabayashi K, Okamoto M, Koshiba T, Kamiya Y, Nambara E. Genome-wide profiling of stored mRNA in *Arabidopsis thaliana* seed germination: epigenetic and genetic regulation of transcription in seed. Plant J. 2005;41: 697–709. doi:10.1111/j.1365-313X.2005.02337.x

444. Winter D, Vinegar B, Nahal H, Ammar R, Wilson GV, Provart NJ. An "Electronic Fluorescent Pictograph" Browser for Exploring and Analyzing Large-Scale Biological Data Sets. PLoS ONE. 2007;2. doi:10.1371/journal.pone.0000718

445. Honaas LA, Wafula EK, Wickett NJ, Der JP, Zhang Y, Edger PP, et al. Selecting Superior *De Novo* Transcriptome Assemblies: Lessons Learned by Leveraging the Best Plant Genome. PLOS ONE. 2016;11: e0146062. doi:10.1371/journal.pone.0146062

446. Wang S, Gribskov M. Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis. Bioinformatics. 2017;33: 327–333. doi:10.1093/bioinformatics/btw625

447. Berget SM, Moore C, Sharp PA. Spliced segments at the 5' terminus of adenovirus 2 late mRNA. Proc Natl Acad Sci U S A. 1977;74: 3171–3175.

448. Gilbert W. The Exon Theory of Genes. Cold Spring Harb Symp Quant Biol. 1987;52: 901–905. doi:10.1101/SQB.1987.052.01.098

449. Koonin EV, Senkevich TG, Dolja VV. The ancient Virus World and evolution of cells. Biol Direct. 2006;1: 29. doi:10.1186/1745-6150-1-29

450. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biol Direct. 2012;7: 11. doi:10.1186/1745-6150-7-11

451. Wahl MC, Will CL, Lührmann R. The Spliceosome: Design Principles of a Dynamic RNP Machine. Cell. 2009;136: 701–718. doi:10.1016/j.cell.2009.02.009

452. Papasaikas P, Valcárcel J. The Spliceosome: The Ultimate RNA Chaperone and Sculptor. Trends Biochem Sci. 2016;41: 33–45. doi:10.1016/j.tibs.2015.11.003

453. Turunen JJ, Niemelä EH, Verma B, Frilander MJ. The significant other: splicing by the minor spliceosome. Wiley Interdiscip Rev RNA. 2013;4: 61–76. doi:10.1002/wrna.1141

454. Qu W, Cingolani P, Zeeberg BR, Ruden DM. A Bioinformatics-Based Alternative mRNA Splicing Code that May Explain Some Disease Mutations Is Conserved in Animals. Front Genet. 2017;8. doi:10.3389/fgene.2017.00038

455. Sasaki-Haraguchi N, Shimada MK, Taniguchi I, Ohno M, Mayeda A. Mechanistic insights into human pre-mRNA splicing of human ultra-short introns: Potential unusual mechanism identifies G-rich introns. Biochem Biophys Res Commun. 2012;423: 289–294. doi:10.1016/j.bbrc.2012.05.112

456. Piovesan A, Caracausi M, Ricci M, Strippoli P, Vitale L, Pelleri MC. Identification of minimal eukaryotic introns through GeneBase, a user-friendly tool for parsing the NCBI Gene databank. DNA Res Int J Rapid Publ Rep Genes Genomes. 2015;22: 495–503. doi:10.1093/dnares/dsv028

457. Abebrese EL, Ali SH, Arnold ZR, Andrews VM, Armstrong K, Burns L, et al. Identification of human short introns. PLOS ONE. 2017;12: e0175393. doi:10.1371/journal.pone.0175393

458. Lewandowska D, Simpson CG, Clark GP, Jennings NS, Barciszewska-Pacak M, Lin C-F, et al. Determinants of Plant U12-Dependent Intron Splicing Efficiency. Plant Cell. 2004;16: 1340–1352. doi:10.1105/tpc.020743

459. Wang G-S, Cooper TA. Splicing in disease: disruption of the splicing code and the decoding machinery. Nat Rev Genet. 2007;8: 749–761. doi:10.1038/nrg2164

460. Will CL, Lührmann R. Spliceosome Structure and Function. Cold Spring Harb Perspect Biol. 2011;3: a003707. doi:10.1101/cshperspect.a003707

461. Jacob M, Gallinaro H. The 5' splice site: phylogenetic evolution and variable geometry of association with U1RNA. Nucleic Acids Res. 1989;17: 2159–2180.

462. Deslandes L, Pileur F, Liaubet L, Camut S, Beynon J, Arlat M, et al. Identification and Mapping of *RRS1*, a Single Recessive Locus in *Arabidopsis thaliana* that Confers Resistance to *Ralstonia solanacearum*. In: Prior P, Allen C, Elphinstone J, editors. Bacterial Wilt Disease: Molecular and Ecological Aspects. Berlin, Heidelberg: Springer Berlin Heidelberg; 1998. pp. 250–254. doi:10.1007/978-3-662-03592-4_36

463. Niu X, Luo D, Gao S, Ren G, Chang L, Zhou Y, et al. A conserved unusual posttranscriptional processing mediated by short, direct repeated (SDR) sequences in plants. J Genet Genomics Yi Chuan Xue Bao. 2010;37: 85–99. doi:10.1016/S1673-8527(09)60028-X

464. Brendel V, Xing L, Zhu W. Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. Bioinforma Oxf Engl. 2004;20: 1157–1169. doi:10.1093/bioinformatics/bth058

465. Sparks ME, Brendel V. Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. Bioinforma Oxf Engl. 2005;21 Suppl 3: iii20-30. doi:10.1093/bioinformatics/bti1205

466. Brent MR, Guigó R. Recent advances in gene structure prediction. Curr Opin Struct Biol. 2004;14: 264–272. doi:10.1016/j.sbi.2004.05.007

467. Huang Y, Chen S-Y, Deng F. Well-characterized sequence features of eukaryote genomes and implications for *ab initio* gene prediction. Comput Struct Biotechnol J. 2016;14: 298–303. doi:10.1016/j.csbj.2016.07.002

468. Shen J, Araki H, Chen L, Chen J-Q, Tian D. Unique Evolutionary Mechanism in R-Genes Under the Presence/Absence Polymorphism in *Arabidopsis thaliana*. Genetics. 2006;172: 1243–1250. doi:10.1534/genetics.105.047290

469. Dietrich RC, Incorvaia R, Padgett RA. Terminal intron dinucleotide sequences do not distinguish between U2- and U12-dependent introns. Mol Cell. 1997;1: 151–160.

470. Sharp PA, Burge CB. Classification of introns: U2-type or U12-type. Cell. 1997;91: 875–879.

471. Mühlemann O, Kreivi JP, Akusjärvi G. Enhanced splicing of nonconsensus 3' splice sites late during adenovirus infection. J Virol. 1995;69: 7324–7327.

472. Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA. Features of *Arabidopsis* Genes and Genome Discovered using Full-length cDNAs. Plant Mol Biol. 2006;60: 69–85. doi:10.1007/s11103-005-2564-9

473. Mayjonade B, Gouzy J, Donnadieu C, Pouilly N, Marande W, Callot C, et al. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. BioTechniques. 2016;61: 203–205. doi:10.2144/000114460

474. Workman R, Fedak R, Kilburn D, Hao S, Liu K, Timp W. High Molecular Weight DNA Extraction from Recalcitrant Plant Species for Third Generation Sequencing. 2018; Available: https://www.nature.com/protocolexchange/protocols/6785

475. Crevillén P, Sonmez C, Wu Z, Dean C. A gene loop containing the floral repressor *FLC* is disrupted in the early phase of vernalization. EMBO J. 2013;32: 140–148. doi:10.1038/emboj.2012.324

476. Ariel F, Jegu T, Latrasse D, Romero-Barrios N, Christ A, Benhamed M, et al. Noncoding Transcription by Alternative RNA Polymerases Dynamically Regulates an Auxin-Driven Chromatin Loop. Mol Cell. 2014;55: 383–396. doi:10.1016/j.molcel.2014.06.011

477. Liu C, Weigel D. Chromatin in 3D: progress and prospects for plants. Genome Biol. 2015;16. doi:10.1186/s13059-015-0738-6

478. Liu C, Wang C, Wang G, Becker C, Zaidem M, Weigel D. Genome-wide analysis of chromatin packing in *Arabidopsis thaliana* at single-gene resolution. Genome Res. 2016;26: 1057–1068. doi:10.1101/gr.204032.116

479. Rosin FM, Watanabe N, Cacas J-L, Kato N, Arroyo JM, Fang Y, et al. Genome-wide transposon tagging reveals location-dependent effects on transcription and chromatin organization in *Arabidopsis*. Plant J. 2008;55: 514–525. doi:10.1111/j.1365-313X.2008.03517.x

480. Varshney RK, Grosse I, Hähnel U, Siefken R, Prasad M, Stein N, et al. Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome. Theor Appl Genet. 2006;113: 239. doi:10.1007/s00122-006-0289-z

481. Roa F, Guerra M. Non-Random Distribution of 5S rDNA Sites and Its Association with 45S rDNA in Plant Chromosomes. Cytogenet Genome Res. 2015;146: 243–249. doi:10.1159/000440930

482. Feng S, Cokus SJ, Schubert V, Zhai J, Pellegrini M, Jacobsen SE. Genome-wide Hi-C analyses in wild type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. Mol Cell. 2014;55: 694–707. doi:10.1016/j.molcel.2014.07.008

483. Grob S, Schmid MW, Grossniklaus U. Hi-C analysis in *Arabidopsis* identifies the KNOT, a structure with similarities to the flamenco locus of *Drosophila*. Mol Cell. 2014;55: 678–693. doi:10.1016/j.molcel.2014.07.009

484. Wang C, Liu C, Roqueiro D, Grimm D, Schwab R, Becker C, et al. Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. Genome Res. 2015;25: 246–256. doi:10.1101/gr.170332.113

485. Sutherland H, Bickmore WA. Transcription factories: gene expression in unions? Nat Rev Genet. 2009;10: 457–466. doi:10.1038/nrg2592

486. Fransz P, de Jong JH, Lysak M, Castiglione MR, Schubert I. Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. Proc Natl Acad Sci U S A. 2002;99: 14584–14589. doi:10.1073/pnas.212325299

487. Tao J-F, Zhou J-Z, Xie T, Wang X-T, Yang Q-Y, Zhang H-Y. Influence of Chromatin 3D Organization on Structural Variations of the *Arabidopsis thaliana* Genome. Mol Plant. 2017;10: 340–344. doi:10.1016/j.molp.2016.09.015

488. Choi Y, Gehring M, Johnson L, Hannon M, Harada JJ, Goldberg RB, et al. DEMETER, a DNA Glycosylase Domain Protein, Is Required for Endosperm Gene Imprinting and Seed Viability in *Arabidopsis*. Cell. 2002;110: 33–42. doi:10.1016/S0092-8674(02)00807-3

489. Dubin MJ, Zhang P, Meng D, Remigereau M-S, Osborne EJ, Casale FP, et al. DNA methylation in *Arabidopsis* has a genetic basis and shows evidence of local adaptation. Elife. 2015;4: e05255.

490. Debladis E, Llauro C, Carpentier M-C, Mirouze M, Panaud O. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. BMC Genomics. 2017;18: 537. doi:10.1186/s12864-017-3753-z

491. Law JA, Ausin I, Johnson LM, Vashisht AA, Zhu J-K, Wohlschlegel JA, et al. A Protein Complex Required for Polymerase V Transcripts and RNA- Directed DNA Methylation in *Arabidopsis*. Curr Biol. 2010;20: 951–956. doi:10.1016/j.cub.2010.03.062

492. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11: 204–220. doi:10.1038/nrg2719

493. Wang X, Weigel D, Smith LM. Transposon Variants and Their Effects on Gene Expression in *Arabidopsis*. PLOS Genet. 2013;9: e1003255. doi:10.1371/journal.pgen.1003255

494. Le TN, Miyazaki Y, Takuno S, Saze H. Epigenetic regulation of intragenic transposable elements impacts gene transcription in *Arabidopsis thaliana*. Nucleic Acids Res. 2015;43: 3911–3921. doi:10.1093/nar/gkv258

495. Bolisetty MT, Rajadinakaran G, Graveley BR. Determining exon connectivity in complex mRNAs by nanopore sequencing. Genome Biol. 2015;16: 204. doi:10.1186/s13059-015-0777-z

496. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10: 1093–1095. doi:10.1038/nmeth.2645

497. Saliba A-E, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42: 8845–8860. doi:10.1093/nar/gku555

498. Efroni I, Birnbaum KD. The potential of single-cell profiling in plants. Genome Biol. 2016;17: 65. doi:10.1186/s13059-016-0931-2

499. Yuan Y, Lee H, Hu H, Scheben A, Edwards D. Single-Cell Genomic Analysis in Plants. Genes. 2018;9. doi:10.3390/genes9010050

500. Iacono G, Mereu E, Guillaumet-Adkins A, Corominas R, Cuscó I, Rodríguez-Esteban G, et al. bigSCale: an analytical framework for big-scale single-cell data. Genome Res. 2018;28: 878–890. doi:10.1101/gr.230771.117

501. Bennett MD, Leitch IJ, Price HJ, Johnston JS. Comparisons with *Caenorhabditis* (∼100 Mb) and *Drosophila* (∼175 Mb) Using Flow Cytometry Show Genome Size in *Arabidopsis* to be ∼157 Mb and thus ∼25 % Larger than the *Arabidopsis* Genome Initiative Estimate of ∼125 Mb. Ann Bot. 2003;91: 547–557. doi:10.1093/aob/mcg057

502. Woo HR, Dittmer TA, Richards EJ. Three SRA-Domain Methylcytosine-Binding Proteins Cooperate to Maintain Global CpG Methylation and Epigenetic Silencing in *Arabidopsis*. PLOS Genet. 2008;4: e1000156. doi:10.1371/journal.pgen.1000156

503. Woo HR, Richards EJ. Natural variation in DNA methylation in ribosomal RNA genes of *Arabidopsis thaliana*. BMC Plant Biol. 2008;8: 92. doi:10.1186/1471-2229-8-92

504. Pontes O, Lawrence RJ, Silva M, Preuss S, Costa-Nunes P, Earley K, et al. Postembryonic Establishment of Megabase-Scale Gene Silencing in Nucleolar Dominance. PLOS ONE. 2007;2: e1157. doi:10.1371/journal.pone.0001157

505. Earley KW, Pontvianne F, Wierzbicki AT, Blevins T, Tucker S, Costa-Nunes P, et al. Mechanisms of HDA6-mediated rRNA gene silencing: suppression of intergenic Pol II transcription and differential effects on maintenance versus siRNA-directed cytosine methylation. Genes Dev. 2010;24: 1119–1132. doi:10.1101/gad.1914110

506. Pontvianne F, Abou-Ellail M, Douet J, Comella P, Matia I, Chandrasekhara C, et al. Nucleolin Is Required for DNA Methylation State and the Expression of rRNA Gene Variants in *Arabidopsis thaliana*. PLOS Genet. 2010;6: e1001225. doi:10.1371/journal.pgen.1001225

507. Earley K, Lawrence RJ, Pontes O, Reuther R, Enciso AJ, Silva M, et al. Erasure of histone acetylation by *Arabidopsis* HDA6 mediates large-scale gene silencing in nucleolar dominance. Genes Dev. 2006;20: 1283–1293. doi:10.1101/gad.1417706

508. Pontvianne F, Blevins T, Chandrasekhara C, Feng W, Stroud H, Jacobsen SE, et al. Histone methyltransferases regulating rRNA gene dose and dosage control in *Arabidopsis*. Genes Dev. 2012;26: 945–957. doi:10.1101/gad.182865.111

509. Jacob Y, Stroud H, LeBlanc C, Feng S, Zhuo L, Caro E, et al. Regulation of heterochromatic DNA replication by histone H3 lysine 27 methyltransferases. Nature. 2010;466: 987–991. doi:10.1038/nature09290

510. Caspi A, Pachter L. Identification of transposable elements using multiple alignments of related genomes. Genome Res. 2006;16: 260–270. doi:10.1101/gr.4361206

511. Kleinboelting N, Huep G, Appelhagen I, Viehoever P, Li Y, Weisshaar B. The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism. Mol Plant. 2015;8: 1651–1664. doi:10.1016/j.molp.2015.08.011

512. Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H. The size and sequence organization of the centromeric region of *Arabidopsis thaliana* chromosome 5. DNA Res Int J Rapid Publ Rep Genes Genomes. 2000;7: 315–321.

513. Jiang J, Birchler JA, Parrott WA, Dawe RK. A molecular view of plant centromeres. Trends Plant Sci. 2003;8: 570–575.

514. Tsukahara S, Kawabe A, Kobayashi A, Ito T, Aizu T, Shin-i T, et al. Centromere-targeted *de novo* integrations of an LTR retrotransposon of *Arabidopsis lyrata*. Genes Dev. 2012; doi:10.1101/gad.183871.111

515. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8: 973–982. doi:10.1038/nrg2165

516. Wright SI, Lauga B, Charlesworth D. Rates and Patterns of Molecular Evolution in Inbred and Outbred *Arabidopsis*. Mol Biol Evol. 2002;19: 1407–1420. doi:10.1093/oxfordjournals.molbev.a004204

517. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, DRABEK J, et al. Evolution of Genome Size in Brassicaceae. Ann Bot. 2005;95: 229–235. doi:10.1093/aob/mci016

518. Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 2009;19: 1419–1428. doi:10.1101/gr.091678.109

519. Devos KM, Brown JKM, Bennetzen JL. Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in *Arabidopsis*. Genome Res. 2002;12: 1075–1079. doi:10.1101/gr.132102

520. Wright SI, Ågren JA. Sizing up *Arabidopsis* genome evolution. Heredity. 2011;107: 509–510. doi:10.1038/hdy.2011.47

521. Bennetzen JL. Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev. 2005;15: 621–627. doi:10.1016/j.gde.2005.09.010

522. Lynch M, Marinov GK. The bioenergetic costs of a gene. Proc Natl Acad Sci U S A. 2015;112: 15690–15695. doi:10.1073/pnas.1514974112

523. Orel N, Puchta H. Differences in the processing of DNA ends in *Arabidopsis thaliana* and tobacco: possible implications for genome evolution. Plant Mol Biol. 2003;51: 523–531.

524. Concia L, Brooks AM, Wheeler E, Zynda GJ, Wear EE, LeBlanc C, et al. Genome-Wide Analysis of the *Arabidopsis* Replication Timing Program. Plant Physiol. 2018;176: 2166–2185. doi:10.1104/pp.17.01537

525. Fawcett JA, Rouzé P, Van de Peer Y. Higher Intron Loss Rate in *Arabidopsis thaliana* Than *A. lyrata* Is Consistent with Stronger Selection for a Smaller Genome. Mol Biol Evol. 2012;29: 849–859. doi:10.1093/molbev/msr254

526. Lynch M. The origins of eukaryotic gene structure. Mol Biol Evol. 2006;23: 450–468. doi:10.1093/molbev/msj050

527. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet. 2013;45: 891–898. doi:10.1038/ng.2684

528. Soria-Hernanz DF, Fiz-Palacios O, Braverman JM, Hamilton MB. Reconsidering the generation time hypothesis based on nuclear ribosomal ITS sequence comparisons in annual and perennial angiosperms. BMC Evol Biol. 2008;8: 344. doi:10.1186/1471-2148-8-344

529. Peterson SN, Fraser CM. The complexity of simplicity. Genome Biol. 2001;2: comment2002.1. doi:10.1186/gb-2001-2-2-comment2002

530. Griffiths-Jones S. Annotating non-coding RNAs with Rfam. Curr Protoc Bioinforma. 2005;Chapter 12: Unit 12.5. doi:10.1002/0471250953.bi1205s9

531. Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics. 2001;2: 8. doi:10.1186/1471-2105-2-8

532. Holmes I. Accelerated probabilistic inference of RNA structure evolution. BMC Bioinformatics. 2005;6: 73. doi:10.1186/1471-2105-6-73

533. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res. 2011;39: D152–D157. doi:10.1093/nar/gkq1027

534. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42: D68–D73. doi:10.1093/nar/gkt1181

535. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997;25: 955–964.

536. Shikata H, Hanada K, Ushijima T, Nakashima M, Suzuki Y, Matsushita T. Phytochrome controls alternative splicing to mediate light responses in *Arabidopsis*. Proc Natl Acad Sci. 2014;111: 18781–18786. doi:10.1073/pnas.1407147112

537. Petrillo E, Herz MAG, Fuchs A, Reifer D, Fuller J, Yanovsky MJ, et al. A chloroplast retrograde signal regulates nuclear alternative splicing. Science. 2014;344: 427–430. doi:10.1126/science.1250322

538. Hartmann L, Drewe-Boß P, Wießner T, Wagner G, Geue S, Lee H-C, et al. Alternative Splicing Substantially Diversifies the Transcriptome during Early Photomorphogenesis and Correlates with the Energy Availability in Arabidopsis. Plant Cell. 2016;28: 2715–2734. doi:10.1105/tpc.16.00508

539. Wang B-B, Brendel V. Genomewide comparative analysis of alternative splicing in plants. Proc Natl Acad Sci. 2006;103: 7175–7180. doi:10.1073/pnas.0602039103

540. Kiegle EA, Garden A, Lacchini E, Kater MM. A Genomic View of Alternative Splicing of Long Non-coding RNAs during Rice Seed Development Reveals Extensive Splicing and lncRNA Gene Families. Front Plant Sci. 2018;9. doi:10.3389/fpls.2018.00115

541. Szakonyi D, Duque P. Alternative Splicing as a Regulator of Early Plant Development. Front Plant Sci. 2018;9. doi:10.3389/fpls.2018.01174

542. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44: D733–D745. doi:10.1093/nar/gkv1189

543. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012;40: D1178–D1186. doi:10.1093/nar/gkr944

544. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14: R36. doi:10.1186/gb-2013-14-4-r36

545. Fournier-Level A, Korte A, Cooper MD, Nordborg M, Schmitt J, Wilczek AM. A map of local adaptation in *Arabidopsis thaliana*. Science. 2011;334: 86–89. doi:10.1126/science.1209271

546. Provart NJ, Alonso J, Assmann SM, Bergmann D, Brady SM, Brkljacic J, et al. 50 years of *Arabidopsis* research: highlights and future directions. New Phytol. 2016;209: 921–944. doi:10.1111/nph.13687

547. Luo M-C, Gu YQ, Puiu D, Wang H, Twardziok SO, Deal KR, et al. Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. Nature. 2017;551: 498–502. doi:10.1038/nature24486

548. Borah P, Sharma E, Kaur A, Chandel G, Mohapatra T, Kapoor S, et al. Analysis of drought-responsive signalling network in two contrasting rice cultivars using transcriptome-based approach. Sci Rep. 2017;7: 42131. doi:10.1038/srep42131

549. Matus JT, Aquea F, Arce-Johnson P. Analysis of the grape MYB R2R3 subfamily reveals expanded wine quality-related clades and conserved gene structure organization across *Vitis* and *Arabidopsis* genomes. BMC Plant Biol. 2008;8: 83. doi:10.1186/1471-2229-8-83

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

551. Brzezinka K, Altmann S, Czesnick H, Nicolas P, Gorka M, Benke E, et al. *Arabidopsis* FORGETTER1 mediates stress-induced chromatin memory through nucleosome remodeling. Weigel D, editor. eLife. 2016;5: e17061. doi:10.7554/eLife.17061

5 Acknowledgements

Many thanks to Prof. Dr. Bernd Weisshaar for providing me with great resources and funding for the *A. thaliana* Nd-1 work as well as general supervision of my PhD. I am also very thankful for the scientific discussions we had during the last years and everything I learned during this time. Many thanks to Prof. Dr. Jörn Kalinowski for co-supervising my PhD and for the interesting discussions.

I would like to thank Dr. Daniela Holtgräwe for great and continuous support since I joined the group. Very helpful advices allowed me to achieve a lot during the last years. Our numerous discussions were crucial for my scientific research and decisions. I also appreciated the cheering up and keeping me motivated during critical times.

Special thanks to Dr. Samuel Brockington for a great supervision during my time in Cambridge including exciting discussions and insightful advices.

My family and friends were a crucial support during the last years. I would like to thank my parents for their constant support during my whole education. Special thanks to Hanna and Nadja for their help and advices especially during the last months.

This work would have been very difficult without constant support from the Bioinformatics Resource Facility team of the CeBiTec. Therefore, I would like to thank the whole team for great technical support.

I am thankful for support from many members of the Chair of Genome Research / Genetics and Genomics of Plants and for the nice time I experienced in this group. Special thanks to Willy and Thomas for sharing optimized protocols and scripts, respectively.

I thank all members of the Brockington group for great support, inspiring discussions, and a nice time in Cambridge.

My participation in events of the DiDy graduate school was very inspiring. I would like to thank all DiDy members for constructive discussions and a great time.

6 Supplements

Publication 1: A *De Novo* Genome Sequence Assembly of the *Arabidopsis thaliana* Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny

Publication 2: Consideration of non-canonical splice sites improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence

Manuscript 1: A Chromosome-level Sequence Assembly Reveals the Structure of the *Arabidopsis thaliana* Nd-1 Genome and its Gene Set

Publication 3: Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes