De novo Wild Daffodil genome assembly to make it flower again

Leandra Vermue

2025, Period 5, 6 and 1

Laboratory of Genetics

Droevendaalsesteeg 1 – 6708 PB Wageningen - The Netherlands

WAGENINGEN
UNIVERSITY & RESEARCH

De novo Wild Daffodil genome assembly to make it flower again

Name student: Leandra
MSc study programme: Bioinformatics
Registration number: 1048003
Course code: GEN80436
Period: 5, 6, 1
Date: 20-10-25
Supervisors: Joost Keurentjes, René Boesten, Rens Holmer, Frank Becker
Examiners: Joost Keurentjes, René Boesten, Rens Holmer

Abstract

The wild daffodil is disappearing in Drenthe and Overijssel, resulting in very small populations. As a consequence, it is expected that the genetic diversity is very low in the wild daffodil populations. However, the current state of genetic diversity is not known, which makes developing conservation strategies harder. To gain insights into the genetic diversity of the wild daffodil populations, wild daffodil individuals can be compared to a reference genome. Yet, a reference genome was not available for the wild daffodil or a related species. Here we present a near-complete reference genome for the wild daffodil, which we used to identify genetic variants in another wild daffodil individual. An estimated genome size of 13,6 Gb was found, confirming prior expectations that the wild daffodil genome is large. The reference genome has a size of 10,8 Gb, only 8,9% of the BUSCO genes are missing, and the assembly consists of 92% of repetitive elements. Furthermore, the reference genome was used for detecting SNPs in another wild daffodil individual using ddRAD sequencing. These results present a reference genome for the wild daffodil, providing a foundation for future genomic studies. Using this genome, the first steps have been taken towards detecting genetic diversity in wild daffodil populations. It is anticipated that these results will enable future studies to determine the genetic diversity within and between wild daffodil populations. This knowledge will provide the basis for designing conservation strategies that ultimately will restore a resilient and healthy wild daffodil population.

Contents

Introduction

Biodiversity is declining globally; 1 million animal, fungal, and plant species are at risk of extinction in a few decades. International policies, like the United Nations Sustainable Development Goals (SDG) or the Convention on Biological Diversity (CBD), are made to reduce this decline. To assess the current status of a species' extinction risk and to measure progress, Red List assessments are made (Hochkirch *et al.*, 2023). The IUCN (International Union for Conservation of Nature and Natural Resources) Red List of Threatened Species is an objective approach for assessing the conservation status of a species. However, there is a strong taxonomic bias in the IUCN Red List of Threatened Species. More than 80% of the assessed species belong to only a few groups (e.g., birds, mammals, and amphibians), compared to only 10% of the described plants that have been assessed (Hochkirch *et al.*, 2021). Plant diversity conservation needs more attention since plants can provide not only food but also medicines, raw materials, and energy. Additionally, plants also act as genetic resources for plant breeding and biotechnological applications. Besides these tangible services, wild plants also have cultural aspects, like recreation and tourism, and serve as a cultural identity. To conserve and maintain plant diversity, the Red List needs to be expanded (Corlett, 2020).

Narcissus pseudonarcissus ssp. Pseudonarcissus (wild daffodil) is an example of a species that is disappearing in the Netherlands. More specifically, in Drenthe and Overijssel, large populations of wild daffodils used to be present but were lost or reduced over time. This decline is mainly due to frequent mowing along ditch edges and the modernization of gardens, which has reduced suitable habitats for the wild daffodil. The wild daffodil is part of the cultural identity of Drenthe and Overijssel, and both provinces aim to preserve and expand the wild daffodil populations. In the proposal for the Red List of vascular plants in 2000, the wild daffodil was categorized as very rare (van der Meijden *et al.*, 2000). The Red List of vascular plants 2012 did not categorize the wild daffodil because it was grouped with all other *Narcissus pseudonarcissus* subspecies. The subspecies grouped together; *N. pseudonarcissus* was categorized as quite rare (Sparrius *et al.*, 2014), which seems like a misleading improvement of the extinction risk of the wild daffodil. If no conservation actions are taken, the wild daffodil in Drenthe and Overijssel may be lost permanently.

Narcissus pseudonarcissus is a monocot, bulb-growing perennial geophyte (Barrett and Harder, 2005), that grows best in damp, poorly drained soils like along river banks and ditch edges. The reproduction of *Narcissus pseudonarcissus* can be clonal, where a mature bulb produces a daughter bulb. Sexual reproduction also occurs where seeds are produced (Caldwell and Wallace, 1955). Through evolution, the genus *Narcissus* underwent chromosome number changes, structural chromosomal rearrangements, and genome size changes (Marques et al., 2017). The genus's taxonomy is thus complex. Based on visual characteristics alone, it is difficult to reliably determine whether a daffodil is a *Narcissus pseudonarcissus ssp. Pseudonarcissus* or the result of historical hybridization with cultivated varieties. Many *Narcissus* (sub)species and cultivars share very similar morphological traits, which makes it challenging to distinguish between naturally occurring wild populations and those that may have been influenced by human cultivation or past hybridization events (Zonneveld, 2008). *Narcissus* can be classified into subgenus *Hermione* and subgenus *Narcissus.* Subgenus *Hermione* has a haploid chromosome number of x = 5, while subgenus *Narcissus* has a haploid chromosome number of x = 7. Other haploid chromosome numbers like x = 10, 11, or 12 are also observed (Zonneveld, 2008; Sochacki *et al.*, 2022). The most frequent ploidy level for the cultivated *Narcissus* is tetraploid, whereas wild populations of *Narcissus* are often diploid (Pustahija *et al.*, 2024). Despite extensive taxonomic

and morphological research, the genetic composition of *Narcissus* species remains largely unknown. Investigating the genetic composition contributes to understanding the genetic diversity, which can provide insights for developing effective conservation strategies for the wild daffodil (Colling *et al.*, 2010).

Genetic diversity is important for the viability of a plant population. A genetically diverse plant population can adapt to changing environmental conditions. It is more likely that a genetically diverse population contains individuals with traits that are more resilient to adverse circumstances, thereby reducing the risk of population decline or extinction. Sexual reproduction between genetically distinct individuals is important in maintaining or enhancing genetic diversity, as it generates new allele combinations (Kardos *et al.*, 2021; Zerebecki and Hughes, 2025). Currently, the wild daffodil is present in very small populations in Drenthe and Overijssel. Colling et al. (2010) studied 15 populations of *N. pseudonarcissus.* using random amplified polymorphic markers. The population size of the studied populations ranged from 100 individuals to 100.000 individuals. The two smallest populations, consisting of 100 and 175 individuals, showed significantly lower genetic diversity than the 13 larger populations, which included at least 250 individuals. Moreover, genetic differentiation between populations grew as their geographical distance increased. These patterns are expected to be present in the wild daffodil populations in Drenthe and Overijssel as well. The current level of genetic diversity is not known for the wild daffodils growing in Drenthe and Overijssel. To explore genetic diversity, a genetic marker analysis can provide useful information, but it offers limited insights and has a low throughput. Alternatively, sequencing reads of wild daffodil individuals can be compared with a reference genome. This provides a broader and faster insight into the genetic diversity. Currently, a reference genome from a species belonging to the *Narcissus* genus does not exist. To provide a broader understanding of the degree of genetic diversity in wild daffodils, a reference genome is needed.

A reference genome is a contiguous and accurate genome assembly representative of a species and provides annotations for genes, regulatory elements, and other functional features. It serves as the foundation for various -omics studies and evolutionary analysis (Formenti *et al.*, 2022). Reference genomes can be generated using short- or long-read sequencing technologies. Although short reads have a low error rate, they often struggle to assemble repetitive regions and complex genomic structures. Long reads can span repetitive or complex regions, which improves the contiguity of the assembly. However, the error rate of long reads produced by Oxford Nanopore Technologies (ONT) is often higher compared to short reads. The reads can be assembled with the use of different algorithms, which use different approaches like overlap-layout-consensus or de Bruijn Graph (Jung *et al.*, 2019). The quality of a reference genome is assessed by determining the assembly statistics, which include N50, the total size, and BUSCO gene counts. The degree of fragmentation of the assembly influences the type of analyses that can be formed. For studies focused on exploring the genetic diversity, moderate fragmentation is usually acceptable. Even with some breaks in repetitive or complex regions, single-nucleotide variants or indels can still be reliably identified (Olkkonen and Löytynoja, 2023). Therefore, a well-assembled reference genome that is not fully contiguous provides a sufficient foundation for exploring genetic diversity within and between wild daffodil populations.

To identify the most suitable wild daffodil individual for the reference genome, a genetic marker analysis was conducted comparing several daffodil populations from Drenthe. This method involved examining specific regions of DNA that are known to vary between species and populations. The

analysis revealed that the daffodil growing in a small forest called the Kloosterbos most likely represents the true wild daffodil (*Narcissus pseudonarcissus ssp. Pseudonarcissus)* (Klein Gotink, 2023). Based only on morphological traits and the marker analysis, it remains impossible to confirm with complete certainty that the daffodils growing in the Kloosterbos are true wild daffodils. The possibility that the daffodils from the Kloosterbos may have originated from a hybridization event in the distant past cannot be entirely ruled out. Despite the remaining uncertainties regarding the origin of the wild daffodil from the Kloosterbos, it was considered the most suitable candidate for establishing a reference genome for the wild daffodil.

The main aim of this research was to assemble a reference genome of the wild daffodil, which would be suitable for detecting genetic diversity within and between wild daffodil populations. Based on literature, it was known that the *Narcissus* species have large genome sizes, so it was hypothesized that the genome size of the wild daffodil would be around 15 Gb (Zonneveld, 2008). Based on the expectation that the genome size of the wild daffodil would be large, it was hypothesized that a large fraction of the genome would consist of repetitive elements. Additionally, it was hypothesized that the assembled reference genome would not be highly contiguous, yet sufficient for detecting genetic diversity. To test these hypotheses, Oxford Nanopore Technologies sequencing was done on the PromethION. A k-mer count was performed on the sequencing data, revealing an estimated genome size of 13,6 Gb. A reference genome of 10,8 Gb was assembled using Shasta, containing 80% complete BUSCO genes and consisting of 146.889 contigs. The portion of repetitive elements was measured for the reference genome, which was a total of 92%. This reference genome was applied in combination with ddRAD sequencing data, and variant calling with BCFtools yielded SNPs. While these first results show that SNPs can be identified using the reference genome, further improvements or adjustments can be made. The assembly could benefit from filtering out (partial) contigs with a lower coverage, and the detected SNPs could be filtered differently. Ultimately, this research is a step towards exploring the genetic variation of wild daffodil populations, which contributes to providing the necessary information to cross genetically distant wild daffodils to stimulate genetic diversity. Most importantly, this will lead to self-sustainable wild daffodil populations in Drenthe and Overijssel, which can withstand biotic and abiotic stresses.

Methods

Plant material

Leaves from the wild daffodils (*Narcissus pseudonarcissus ssp. Pseudonarcissus)* were sampled in Nijeveen (Dorpsstraat 153) and in the Kloosterbos located near Schoonebeek (52°39'00.8"N 6°51'47.1"E). The entire leaf was cut using scissors and collected with tweezers. Multiple leaves were sampled from each wild daffodil plant. Between each wild daffodil plant, the scissors and tweezers were cleaned using 100% ethanol and dried with tissues. The samples were kept on ice until frozen in liquid nitrogen and stored at -80°C. The commercial cultivar Tête-à-Tête was bought at the store and sampled in the same way, but not stored at -80°C.

Flow cytometry

Leaf material from the wild daffodil sampled in the Kloosterbos and the Daffodil Tête-à-Tête was sent for analysis to Plant Cytometry Services. Flow cytometry was performed to measure the relative genome size of each sample, using *Monstera deliciosa*, *Ophiopogon planiscapus* 'Niger', and *Clivia*

miniata as reference standards. Absolute and relative fluorescence values were calculated for each sample.

DNA isolation & library preparation

DNA from the wild daffodil from the Kloosterbos and Nijeveen was isolated following the protocol described by Driguez et al., (2021) using the QIAGEN DNA Buffer Set and the QIAGEN Genomic-tip 100/G. The following adjustments were made to the protocol; 500 mg, instead of 1 g, of ground leaf powder was resuspended in lysis buffer. The suspension was incubated at 50°C, for 3,5 hours and additionally spun at 300 rpm. The supernatant was not vortexed but poured directly onto the calibrated genomic-tip. The DNA was eluted in 5 mL prewarmed (37°C) buffer QF. The DNA 'jellyfish' was dissolved in 70 µL 10 mM Tris buffer (pH 8) instead of EB buffer.

The extracted DNA from the Kloosterbos was further prepped for Oxford Nanopore Technology (ONT) sequencing. For this, the Circulomics Short Read Eliminator XL kit was used following the manufacturer's protocol to eliminate the short reads from the extracted DNA. The short read eliminated DNA was used for DNA repair and end-prep, followed by the adapter ligation and clean-up (Ligation Sequencing Kit V14 (SQK-LSK114) protocol from Oxford Nanopore Technologies). The Ligation Sequencing Kit V14 from Oxford Nanopore Technologies and the NEBNext® Companion Module v2 for Oxford Nanopore Technologies® Ligation Sequencing were used.

Sequencing and basecalling

Two PromethION flow cells (R10.4.1) were used for sequencing on the PromethION from Oxford Nanopore Technologies. The flow cells were prepared and loaded following the Priming and loading the PromethION Flow Cell protocol (Ligation Sequencing Kit V14 (SQK-LSK114) protocol from Oxford Nanopore Technologies). When the pore availability dropped below 500, the flow cell was washed using the Flow Cell Wash Kit from Oxford NanoPore Technologies following the manufacturer's protocol and loaded with the library. When loading the library for the third time onto the flow cell, it ran for 72 hours. During sequencing, the reads were immediately basecalled with the high accuracy model v4.3.0, 400 bps. Later, the reads were basecalled again using Dorado (0.7.2) basecaller with the model dna_r10.4.1_e8.2_400bps_sup@v5.0.0, the rest was left on default. To remove the reads with an average Phred read quality score below 10, Chopper 0.10.0 (De Coster and Rademakers, 2023) was used, all options were left at default. To assess the quality of the reads, NanoPlot 1.46.1 (De Coster and Rademakers, 2023) was used using the fastq option, the rest was left on default.

K-mer counting

Jellyfish count and histo 2.2.10 (Marçais and Kingsford, 2011) were used to count the k-mers from the ONT reads and make a histogram. A k-mer size of 21 was used, the rest was left on default. GenomeScope 2.0 (Ranallo-Benavidez *et al.*, 2020) was used with ploidy 2 to visualize this histogram, the rest was left on default.

Assembly

Shasta 0.14.0 (Shafin *et al.*, 2020) or Miniasm-0.3 (r179) (Li, 2016) was used to make an assembly. Other assembly algorithms (Flye 2.9.5 (Kolmogorov *et al.*, 2019), Raven 1.8.3 (Vaser and Šikić, 2020)) were tried but failed due to computational reasons. The Nanopore-Plants-Apr2021 configuration was used when creating assemblies with Shasta. The Shasta parameters for an assembly where either left on default or the option of minimum read length was set to 0. For making an assembly with Miniasm, Minimap2 2.28 (Li, 2016) was first used to align the reads with the option Oxford Nanopore all-vs-all

overlap mapping. The parameters for Miniasm and Minimap2 were left at default. Every assembly was analysed with Quast 5.0.2 (Mikheenko *et al.*, 2018) to assess the contig sizes, all options were left at defailt. Busco 5.8.3 (Manni *et al.*, 2021) was used with the liliopsida_odb12 database to check how many conserved genes are present in the assembly, all options were left on default.

Repeats

To assess the repeat content of the assembly, RepeatModeler 2.0.6 (Smit and Hubley) and RepeatMasker 4.1.9 (Smit *et al.*) were used. With RepeatModeler BuildDatabase a database was made, using assembly 5 (Table 3), all options were left at default. After building a database, RepeatModeler was used to discover the repeats in the database. LTRstruct was enabled, and the rest of the options were left at default. After obtaining the repeat library from RepeatModeler, RepeatMasker was used to mask the repeats in the genome and to assess the repeat content of the assembly. RepeatMasker masked the repetitive regions in assembly 6 (Table 3) using the earlier build database, rmblast 2.14.1+ was used as the search engine, rush job was enabled, and the repetitive regions were returned with a lowercase.

ddRAD sequencing

In vitro digestion and sequencing

The reduction of the genome was done with double-digest Restriction Associated DNA sequencing (ddRADseq). The double digest protocol described by Peterson et al. (2012) was carried out in triplicate. Some adjustments were made to the protocol; for the double digest, 500 ng of DNA was used from the wild daffodils collected at Kloosterbos and Nijeveen. 20 units instead of 10 units per restriction enzyme were added. 10x rCutSmart™ Buffer, provided by New England Biolabs was added as appropriate buffer for the restriction enzymes. The enzyme combination SbfI-HF and EcoRI-HF provided by New England Biolabs was used to achieve as much reduction as possible (Peterson *et al.*, 2012). After incubation, from one replicate, 30 µL was taken and divided into three aliquots of 10 µl. Each aliquot was treated separately: one with 0,6X AMPure XP beads, one with 1X AMPure XP beads, and one with 1,5X AMPure XP beads (provided by Beckman Coulter). The eluted DNA from the beads was run on a 1% gel to assess the reduction of the genome. Based on the results, the 0,6x AMPure XP bead treatment was selected for further treatment on the remaining repeats. The supernatant from the 0.6X beads was again treated with 1,5X AMPure XP beads. The DNA from the 1,5x AMPure XP beads was eluted in 15 µl nuclease-free water.

The library preparations were conducted according to the Hackflex protocol (Gaio *et al.*, 2022). The following alterations were made to the protocol. After adding the SDS, the samples were incubated for 10 min at 37°C instead of 15 min at 37°C. After incubation, the PCR plate was placed on a magnet, and after 4 minutes, the beads were washed with washing solution while keeping the PCR plate on the magnet. After washing 22,5 µl PrimeSTAR GXL Premix PCR master mix (Takara), 20 µL nuclease-free water and 2,5 µL oligo's were added to the beads. This was incubated in the T100 Thermal Cycler (Bio-Rad) for 14 cycles. After incubation, the samples were pooled, and 100 µl was taken from the pooled sample for size selection. 160 µL of diluted SPRI beads (109,25 µL SPRI beads + 74,75 µL MilliQ water) was added to the 100 µL pooled sample and incubated for 5 minutes, after which it was placed on the magnet for 5 minutes. 250 µL supernatant was transferred to a new tube. The remaining beads were washed twice with 80% ethanol and eluted in 26 µL MilliQ water; this is size fraction 1 of the sample. To the supernatant, 30 µL undiluted SPRI beads were added, and after incubation as described before, 280 µL supernatant was transferred to a new tube. The remaining beads were washed and

eluted as described earlier; this is size fraction 2 of the sample. To the new supernatant, 112 µL undiluted SPRI beads were added. After incubation as described earlier, all supernatant should be transferred. The remaining beads were washed and eluted as earlier described; this is size fraction 3 of the sample. 2 µL from the eluted DNA from all size fractions was checked on a 2% agarose gel. After inspection, size fraction 2 of the sample ranged around 500 bp and was sent for Illumina sequencing provided by Novogene.

In silico digestion

To get an indication of the fragmentation pattern after the double digest, an *in silico* digestion was carried out. A Python script was developed to mimic the action of the restriction enzymes (supplementary data). The script identified the restriction recognition sites within assembly 6 (Table 3) and removed the restriction recognition site, which resulted in fragmentation of the assembly. The resulting fragment size distribution was summarized in a histogram.

Variant calling

Fastp 1.0.1 was used to assess the quality of the Illumina reads (Chen, 2023).

The WGS/WES Mapping to Variant Calls workflow from Samtools was followed to detect markers. From the major steps, Improvement was not done on the alignments. BWA 0.7.19-r1273 (Li, 2013) and BCFtools 1.22 (Li, 2011) were used. During the workflow, the sequencing data from the Kloosterbos and Nijeveen samples were mapped to assembly 6 (Table 3). First BCFtools mpileup was done with one sample: the alignment data from Nijeveen. The second time, BCFtools mpileup was done with two samples: the BWA alignment data from Nijeveen and the Kloosterbos.

Two different pipelines from Stacks 2.68 (Catchen *et al.*, 2011) were also used to detect markers. Before running the ref_map.pl pipeline, the sequencing data from the Kloosterbos and Nijeveen were mapped to assembly 6 (Table 3) using Bowtie2 2.5.4 (Langmead and Salzberg, 2012). The mapping was done end-to-end, the rest of the options were left at default. The ref_map.pl pipeline was run twice. First, adding one sample: the bowtie2 alignment data from Nijeveen. The second time, the ref_map.pl pipeline was run with two samples: the Bowtie2 alignment data from Neijeveen and the Kloosterbos. All options were left on default. The denovo_map.pl pipeline was used on the sequencing data of Nijeveen and the Kloosterbos, all settings were left at default.

To obtain variant call data, the populations program from Stacks was run on the output of each of the pipelines described above. All settings were left at default, except for the addition of the option to generate a Variant Call Format. To filter the called SNPs VCFtools 0.1.17 (Danecek *et al.*, 2011) and BCFtools view 1.22 (Danecek *et al.*, 2021) were used.

The BAM files generated by mapping the RADseq data to the reference genome using BWA and Bowtie2 were analysed using Samtools stats 1.6 (Danecek *et al.*, 2021). The results were summarized and visualized using MultiQC 1.31 (Ewels *et al.*, 2016).

Results

The genome size of a triploid is 10% bigger than the wild daffodil from the Kloosterbos

To confirm if the wild daffodil from the Kloosterbos is a diploid, flow cytometry was done on leaf material. Leaf material from the Kloosterbos and the commercial cultivar Tête-à-Tête, a known triploid, was used. When the Monstera and Ophiopogon standards were used, the absolute and relative fluorescence values were consistently ~10% higher for Tête-à-Tête than for the Kloosterbos (Table 1). The fluorescence values measured with the standard Clivia did not show a difference in absolute values, but the relative values were 11% higher for Tête-à-Tête than for the Kloosterbos. Across all standards that were measured, the genome size of Tête-à-Tête is ~10% bigger than the genome size of the wild daffodil growing in the Kloosterbos. It was expected that the wild daffodil from the Kloosterbos is a diploid and since Tête-à-Tête is a known triploid; these results do not align with the expected 1.5-fold difference between a diploid and a triploid. Nonetheless, we still proceeded with the wild daffodil from the Kloosterbos for whole-genome sequencing.

Table 1. The flow cytometry data from the Kloosterbos and Tête-à-Tête. The absolute (abs) ratio and the relative (rel) ratio were calculated with 3 different standards.

	Standards								
	Monstera			**Ophiopogon**			**Clivia**		
Sample	**abs**	**rel**	**rel/abs**	**abs**	**rel**	**rel/abs**	**abs**	**rel**	**rel/abs**
Kloosterbos	3,75	4,14	1,10	2,79	2,74	0,98	±1	0,81	-
Tête-à-Tête	4,13	4,63	1,12	3,00	3,06	1,02	±1	0,89	-
Tête-à-Tête/Kloosterbos	1,10	1,12		1,08	1,12		-	1,11	

A total of 166,35 Gb sequencing data was generated

To be able to assemble a reference genome of the wild daffodil, the extracted DNA from the wild daffodil sampled in the Kloosterbos was sequenced using the PromethION. The first flow cell yielded 83,28 Giga bases (Gb) and the second flow cell 83,06 Gb. In total, 11,57 million reads were generated, adding up to 166,34 Gb. The first flow cell yielded an average N50 of 28,70 kb, and the second flow cell yielded an average N50 of 26,10 kb. The overall average N50 of the reads was 27,40 kb (Table 2).

Table 2. Information per sequencing run and the total and average for all sequencing runs.

		Sequenced		
Flow cell	**Run**	**Bases (Gb)**	**Reads (M)**	**N50 (kb)**
1	1	30,35	1,86	28,76
1	2	26,30	1,68	28,07
1	3	26,63	1,78	29,27
1	**Average**	**-**	**-**	**28,70**
1	**Total**	**83.28**	**5,32**	**-**
2	1	33,99	2,44	28,37
2	2	25,90	1,96	25,08
2	3	23,17	1,85	24,86
2	**Average**	**-**	**-**	**26,10**
2	**Total**	**83,06**	**6,25**	**-**
1 + 2	**Average**	**-**	**-**	**27,40**
1 + 2	**Total**	**166,34**	**11,57**	**-**

Improved read quality after basecalling with the super accurate model

The read quality was checked before the reads could be used in further analysis. To assess the read quality, NanoPlot was used. The reads that were basecalled with the high accuracy (HAC) model yielded 162,6 Gb with an average Phred read quality score above 7. The average Phred quality of all reads was 13,9 and the median Phred quality was 15,1. The read quality histogram shows a peak around 15 (Figure 1A). The reads that were basecalled with the super accurate (SUP) model yielded 163,2 Gb with an average Phred read quality score above 10. The average Phred quality score of all reads was 19,4 and the median Phred read quality score was 21,8. The read quality histogram shows a peak between 20 and 25 (Figure 1B). A 39,6% increase in average Phred read quality score was obtained by basecalling with the SUP model compared to the HAC model.

In both scatterplots, it can be observed that the majority of the reads are shorter than 20 kb, and almost all reads are shorter than 40 kb (Figure 1). The shorter reads (< 10 kb) displayed a broad range of Phred quality scores, whereas longer reads tend to cluster at slightly lower than average, but more consistent Phred quality scores.

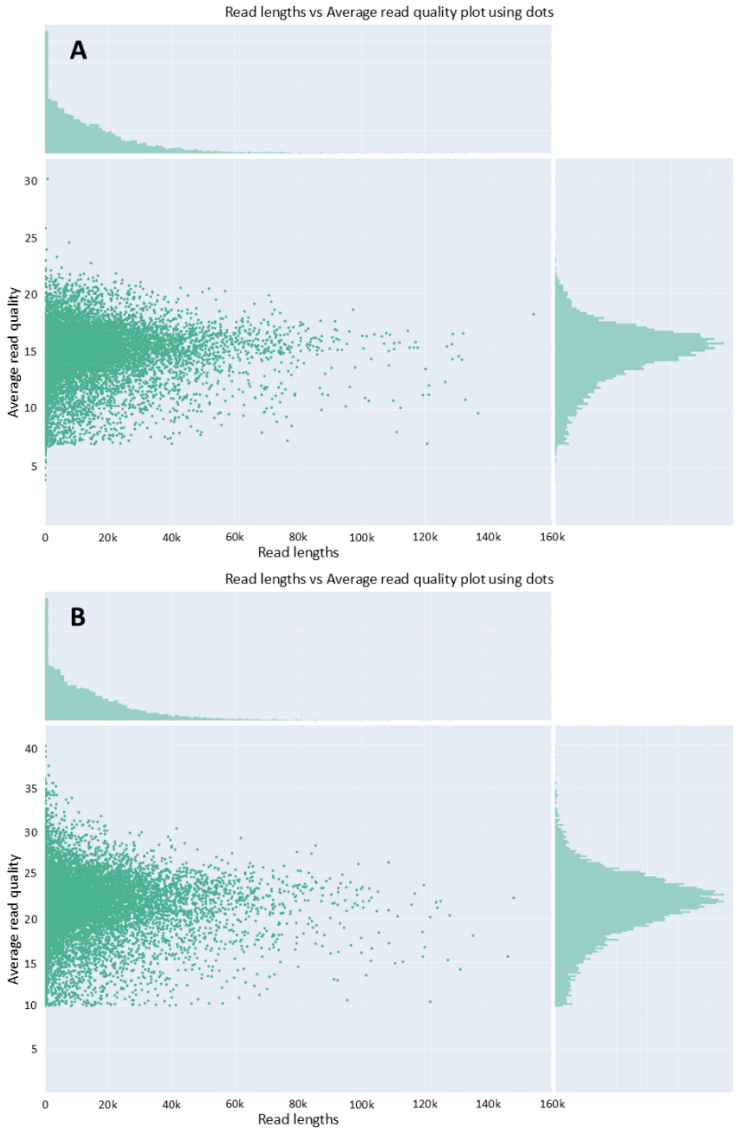

Figure 1. Scatterplot of the average Phred read quality scores plotted against the read length, and a histogram at each side of the scatterplot of the read length and read quality. A) Reads were basecalled with the high accuracy model, and reads with an average quality below 7 were discarded. B) Reads were basecalled with the super accurate model, and reads with an average quality below 10 were discarded.

The estimated genome size of the wild daffodil is 13,6 Gb

To estimate the genome size of the wild daffodil sampled at the Kloosterbos, a k-mer count was done on the sequencing data. The k-mer count was visualized in a histogram using GenomeScope2 (Figure 2). No peak was observed in the k-mer histogram of reads basecalled with the HAC model (Figure 2A). In the k-mer histogram of reads basecalled with the SUP model, a peak was visible (Figure 2B). The peak is present at a coverage of 12. A total of 163,2 Gb SUP basecalled reads were used during k-mer counting. Dividing the total amount of reads by the coverage corresponds to an estimated genome size of 13,6 Gb for the wild daffodil.

For both histograms, the full model from GenomeScope2 did not fit the observed histogram (Figure 2). In one histogram, the error line follows a similar peak to the observed peak (Figure 2B). The estimated length of the genome calculated by the mathematical model also does not correspond with the calculated 13,6 Gb from the observed peak. Since the full model of GenomeScope2 fitted the observed peak poorly, the estimated genome size of the wild daffodil is considered to be 13,6 Gb. The estimated genome size of 13,6 Gb will also be the expected size when the whole genome is assembled.

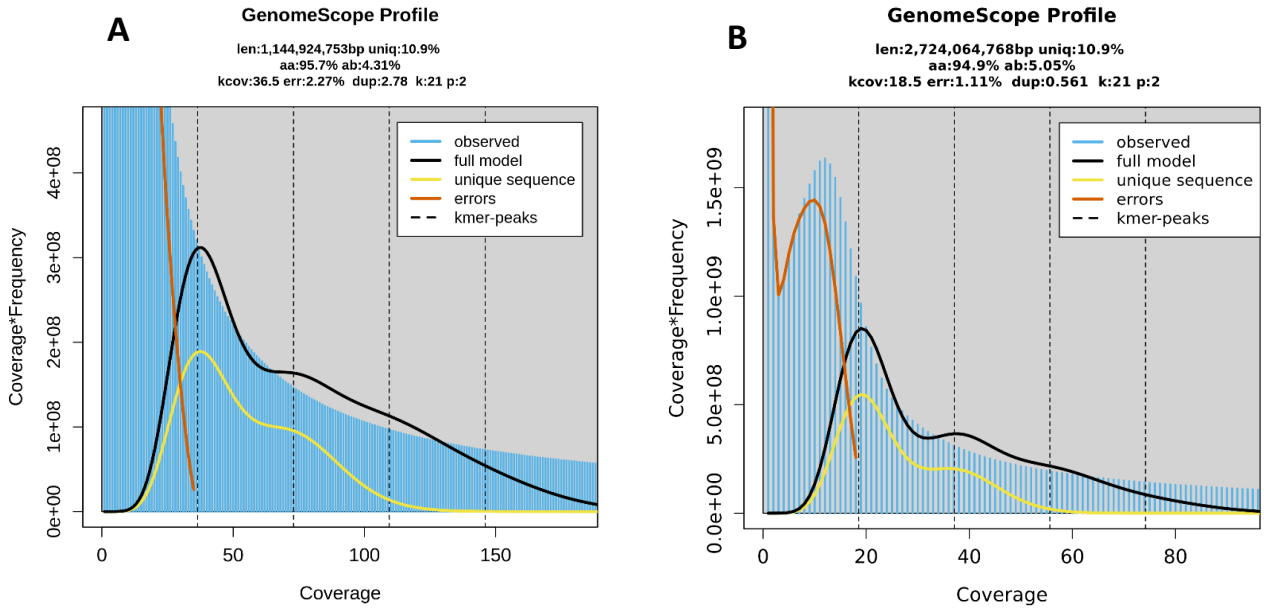

Figure 2. GenomeScope 2.0 transformed linear plot. A) The k-mer count was done on ONT reads that were basecalled with the high accuracy model and had an average read Q-score higher than 7. B) The k-mer count was done on ONT reads that were basecalled with the super accurate model and had an average read Q-score higher than 10.

Multiple assemblies were generated with different algorithms

After obtaining sequencing data from the first flow cell, the first assemblies were made. The aim was to generate a reference genome for the wild daffodil that is as complete as possible and comparable in size to the estimated genome size. To generate the reference genome, multiple assembly algorithms were tested to determine which was most suitable for the dataset. To determine the quality of the assembly, different measurements were done. The total length was assessed to determine if it resembled the estimated genome size. The number of contigs, N50, and L50 were calculated to get an understanding of the fragmentation and contiguity of the assembly. BUSCO genes are highly conserved genes that are typically present in members of a given taxonomic group. Therefore, the BUSCO gene counts can provide an indication of the completeness of the assembly.

The first assemblies were made using Miniasm and Shasta (Assembly 1 and 3, Table 3). Assembly 3, made with Shasta, is approximately three times longer and contains approximately 30 times more BUSCO genes than assembly 1, made with Miniasm. In assembly 3 made with Shasta, 46,3% of the BUSCO genes were missing, which suggested that the coverage was not high enough to assemble the entire genome. To investigate this, assemblies were created using Shasta with 20%, 40%, 60%, and 80% of the sequencing data generated from one flow cell. The total length, max contig size, # contig (> 50.000 bp), and the BUSCO gene count of each assembly were plotted in the graphs (Figure 3). In all the graphs, except the Max contig size graph, a linear trend between 60% and 100% can be observed. This trend suggests that generating additional sequencing data would improve the completeness of the assembly. Therefore, a second flow cell was used to increase the sequencing depth.

After obtaining the sequencing data from the second flow cell, two additional assemblies were created using Miniasm and Shasta (Assembly 2 and 4, Table 3). Assembly 4 made by Shasta outperformed assembly 2 made by Miniasm on every aspect measured. Notably, the extra generated sequencing data from the second flow cell did not improve the assembly made by Miniasm (Assembly 1 and 2, Table 3). Based on these results, only Shasta was used for all subsequent assemblies to achieve further improvements. To assess whether the linear trend observed between 60% and 100% was leveling off (Figure 3), assembly 4 was included in these graphs. Assembly 4 was generated with additional data from the second flow cell, and is therefore represented as the 200% datapoint. The total length, max contig size, # contigs (>50.000 bp), and the BUSCO gene count for assembly 4 were added. Across all graphs, except the max contig size graph, the 200% datapoint falls slightly below the linear trend observed between 60% and 100% of the dataset.

To further improve the assembly, the sequencing data was basecalled again with the super accurate (SUP) model. With the sequencing data basecalled with SUP, two assemblies were made using Shasta (Assembly 5 and 6, Table 3). Assembly 5 was made with the same parameters as assembly 4, both assemblies 4 and 5 showed similar statistics. However, assembly 5 contained 8 fewer BUSCO genes compared to assembly 4, even though the reads used for assembly 5 had a higher average read quality than the reads used for assembly 4 (Figure 1). For assembly 6, the SUP basecalled reads were used with a minimum read length of 0, which was 10.000 in earlier assemblies made by Shasta. Assembly 6 yielded a total length of 10,8 Gb, which is the closest to the estimated 13,6 Gb genome size of the wild daffodil. 8,9% of the BUSCO genes are missing, and 79,7% of the BUSCO genes are complete in assembly 6. Based on the assembly statistics (Table 3), assembly 6 was selected as the reference genome for performing variant calling.

Table 3. Assembly statistics per assembly. The row named Details lists details about the parameters used for the assembly, if sequencing data from one or two flow cells was used, and which model was used for basecalling: high accuracy (HAC) model or super accurate (SUP) model.

Assembly	**1**	**2**	**3**	**4**	**5**	**6**
Algorithm	Miniasm	Miniasm	Shasta	Shasta	Shasta	Shasta
Details	Reads from one flow cell; bascalled with HAC	Reads from two flow cells; basecalled with HAC	Reads from one flow cell; basecalled with HAC; minimum read length 10.000	Reads from two flow cells; basecalled with HAC; minimum read length 10.000	Reads from two flow cells; basecalled with SUP; minimum read length 10.000	Reads from two flow cells; basecalled with SUP; minimum read length 0

Total length	1.223.178.654	525.083.117	3.955.791.121	10.134.886.316	10.270.391.232	10.756.021.779
# Contigs	35.580	20.474	111.854	132.147	130.880	146.889
# Contigs (> 50000bp)	5.418	1.235	26.371	54.115	54.463	52.942
N50	37.244	28.018	64.395	166.770	165.900	187.499
L50	10.915	6.254	18.394	15.889	15.785	14.076
GC-%	41,28	41,55	40,02	40,31	40,35	40,35
Max contig size	257.016	268.567	603279	2.629.279	2.116.727	3.451.291
BUSCO-total	2821	2821	2821	2821	2821	2821
Complete	17 (0,60%)	3 (0,11%)	1016 (36,0%)	2147 (76,1%)	2131 (75,5%)	2247 (79,7%)
Single	17 (0,60%)	2 (0,07%)	935 (33,1%)	1701 (60,3%)	1676 (59,4%)	1774 (62,9%)
Duplicated	0	1 (0,04%)	81 (2,9%)	446 (15,8%)	455 (16,1%)	473 (16,8%)
Fragmented	15 (0,53%)	4 (0,14%)	500 (17,7%)	355 (12,6%)	363 (12,9%)	322 (11,4%)
Missing	2789 (98,9%)	2814 (99,8%)	1305 (46,3%)	319 (11,3%)	327 (11,6%)	252 (8,9%)

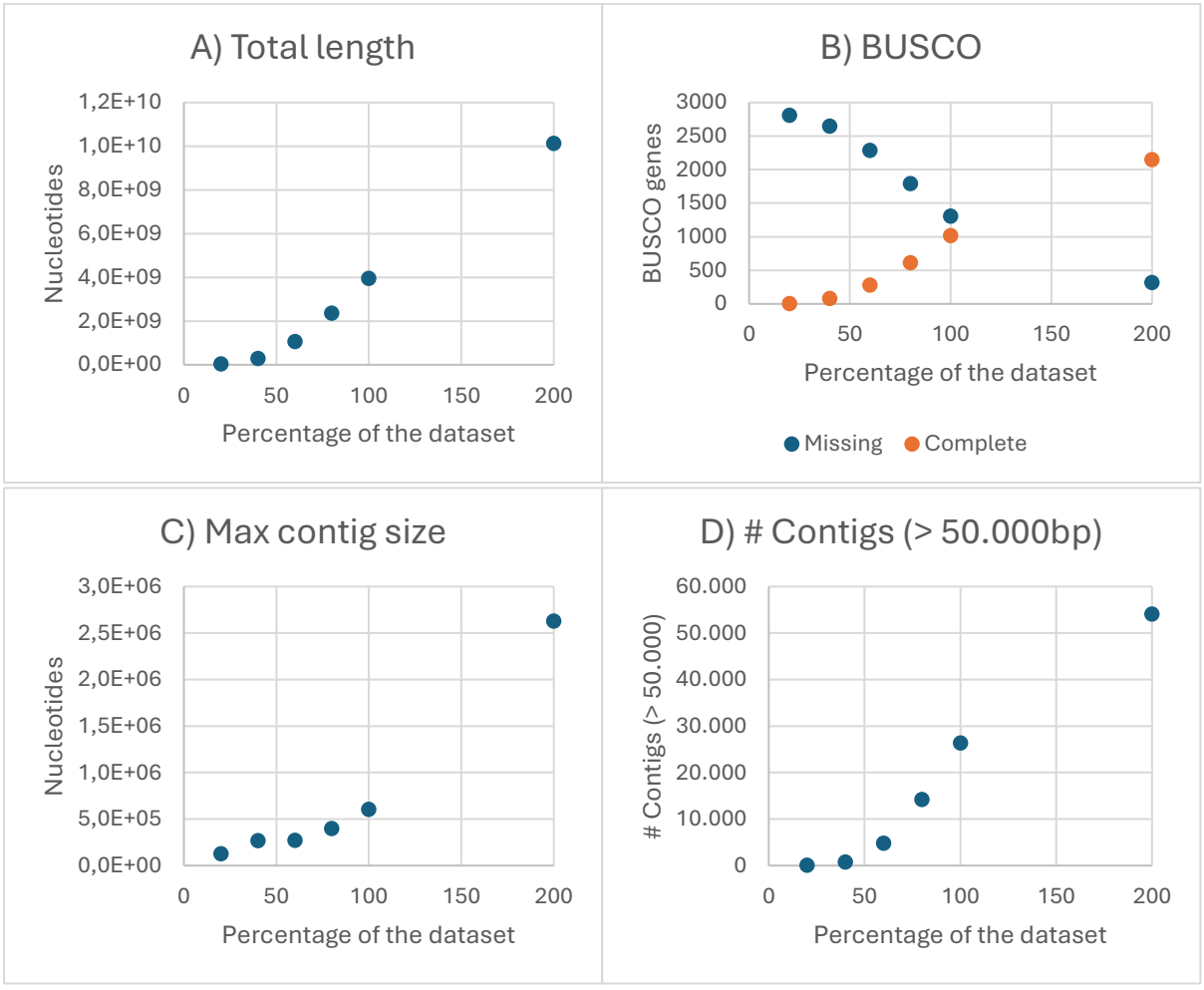

Figure 3. Assembly statistics of Shasta assemblies plotted against the percentage of the dataset used. Four graphs show A) the total length of the assemblies, B) the BUSCO gene count, C) the largest contig in the assemblies, and D) the number of contigs > 50.000 bp. The x-axis represents the percentage of the sequencing dataset used for each assembly. Data points up to 100% correspond to assemblies generated

from the first flow cell, while the 200% data point represents the assembly generated with sequencing data from two flow cells. All assemblies were produced with Shasta using identical parameters. The assembly generated from 100% of the dataset corresponds to assembly 3, and the assembly from 200% of the dataset corresponds to assembly 4 (Table 3).

91,97% of the assembly consists of repetitive elements

To assess the repeat content of the assembly RepeatModeler and RepeatMasker were used. In total, 91,97% of the assembly was recognized as repetitive elements. The majority of the repetitive elements were classified as retro-elements (61,16%) (Figure 4). 29,37% of the recognized repeats could not be classified. The remaining 1,44% of the repetitive elements were classified in very small percentages as DNA transposons, small RNA, simple repeats, and low complexity repeats.

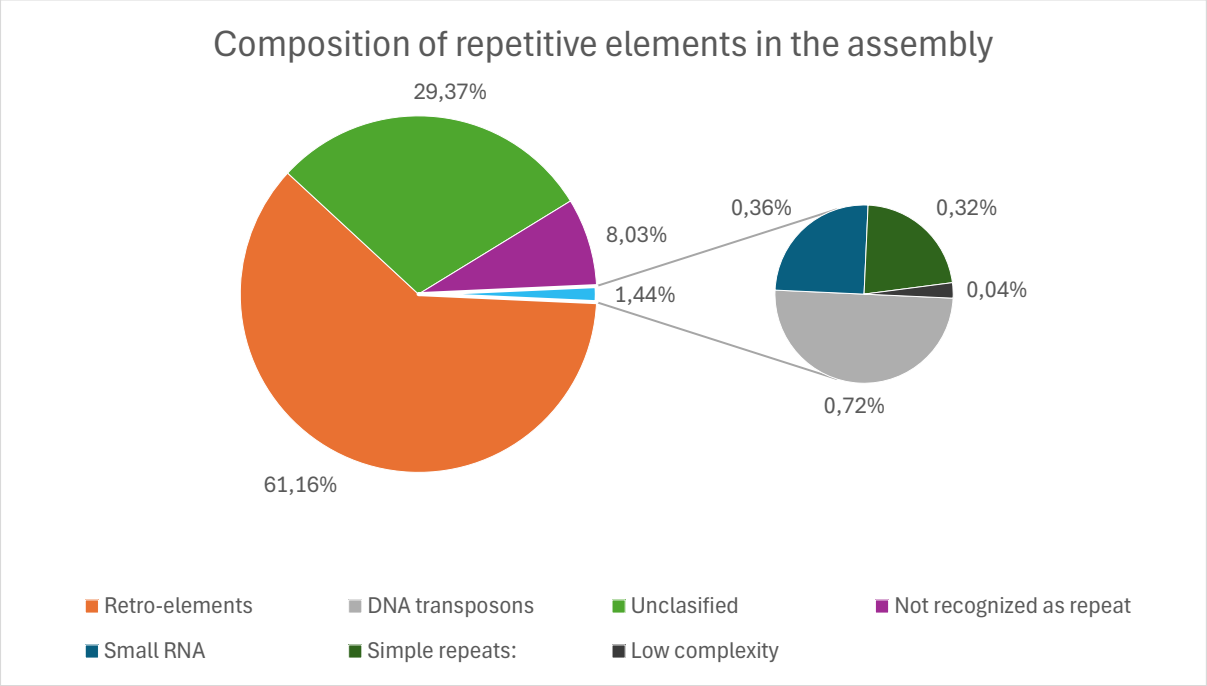

Figure 4. The composition of repetitive elements in assembly 6 (Table 3). In total, 91,97% is recognized as repetitive elements.

Double-digest restriction associated DNA

With a reference genome now available for the wild daffodil, it could be used for variant calling. During variant calling, the sequencing reads of another wild daffodil individual are compared to the reference genome to detect single nucleotide polymorphisms (SNPs). The identified SNPs can be used for detecting genetic diversity in wild daffodils. Whole genome sequencing of another wild daffodil individual would be expensive, therefore, it was chosen to do double digest restriction associated DNA (ddRAD) sequencing. With ddRAD sequencing, a reduced portion of the genome is sequenced, which enables variant calling on the sequenced regions.

In silico double digestion

To predict the fragment size distribution of the genome after double digestion, an *in silico* digestion was performed on assembly 6 of the wild daffodil (Table 3). The fragment size distributions were compared before and after the double digestion (Figure 5). Before digestion, the assembly contained a wide range of fragment sizes, with the majority of the fragments below 250 kb. After digestion, the size distribution shifted towards smaller fragments; the majority of the fragments are below 100 kb. The frequency of the short fragments almost tripled. The zoomed-in size distributions show different patterns before and after digestion (right graphs, Figure 5). Before digestion, the distribution shows a relatively uniform stepwise decrease in fragment counts with increasing fragment size. The zoomed-

in distribution after digestion does not show this stepwise pattern. The distribution is more irregular, with the highest peak not at the smallest fragment size, but at a slightly larger size, followed by an abrupt drop to about half the frequency in the next bin.

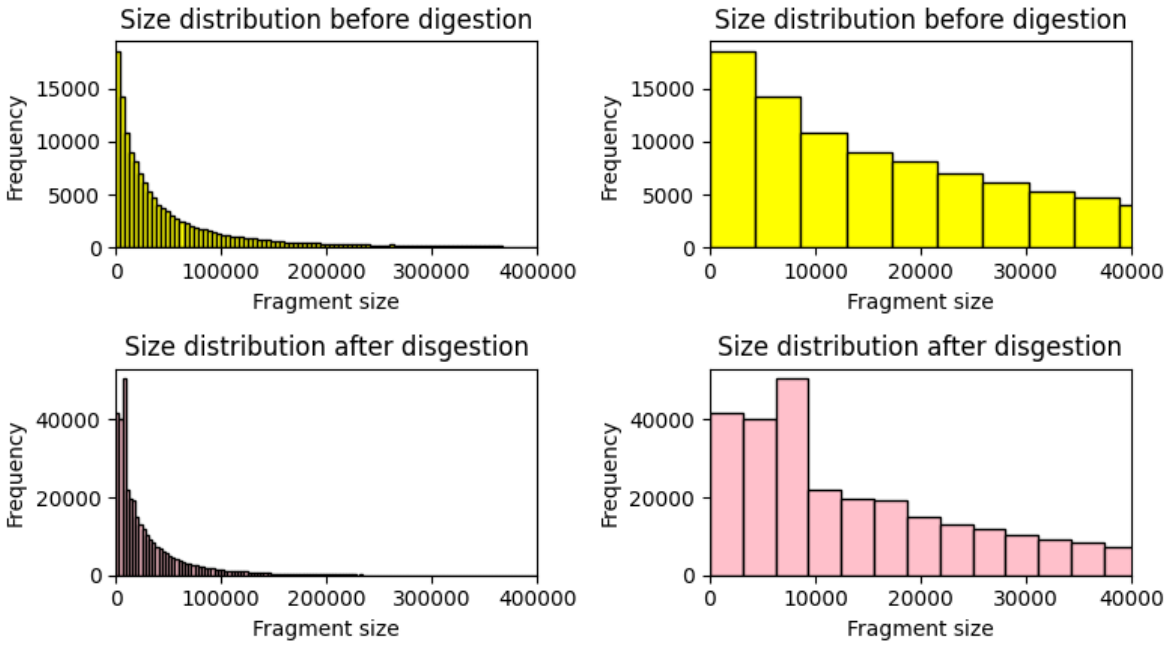

Figure 5. In silico digestion was done on assembly 6 (Table 3). The frequency of the fragment sizes was plotted in the histograms. The graphs on the left display the entire size distribution, while the graphs on the right show a zoomed-in representation (fragment size: 0-40kb).

In vitro double digestion

To ensure that the ddRAD protocol would work and a reduction of the genome was achieved, the digested DNA was treated with different ratios of beads and visualized on a gel (Figure 6A). This was done to evaluate the size of the DNA eluted from the bead ratios. Theoretically, lower bead ratios will elute a lower amount of DNA and longer DNA fragments compared to higher bead ratios. Higher bead ratios will elute a higher amount of DNA and long and shorter DNA fragments, which could be observed on a gel. Overall, no major differences were observed in the amount and fragment size of the DNA eluted from the beads across the different ratios. However, the band on the gel containing DNA eluted from the 0,6X beads appeared slightly less elongated than the bands from the other bead ratios. Based on this observation and how the beads theoretically should work, it was chosen to continue with the supernatant from the 0,6X bead treatment to maximize the genome reduction.

The digested DNA, undigested DNA, eluted DNA from the 0,6X beads, and the supernatant from the 0,6X beads were visualized on a gel (Figure 6B). The undigested DNA shows a high-intensity smear located at the top of the gel, whereas the digested DNA was clearly fragmented as expected after the double digest. No visible DNA bands were detected in either the supernatant of the 0,6X bead treatment or in the DNA eluted from the 0,6X beads. This indicates that the DNA concentration in these samples was below the detection limit of the gel. To obtain sufficient material for sequencing, the supernatant from 0,6X bead treatment was treated with an additional 1,5X bead treatment.

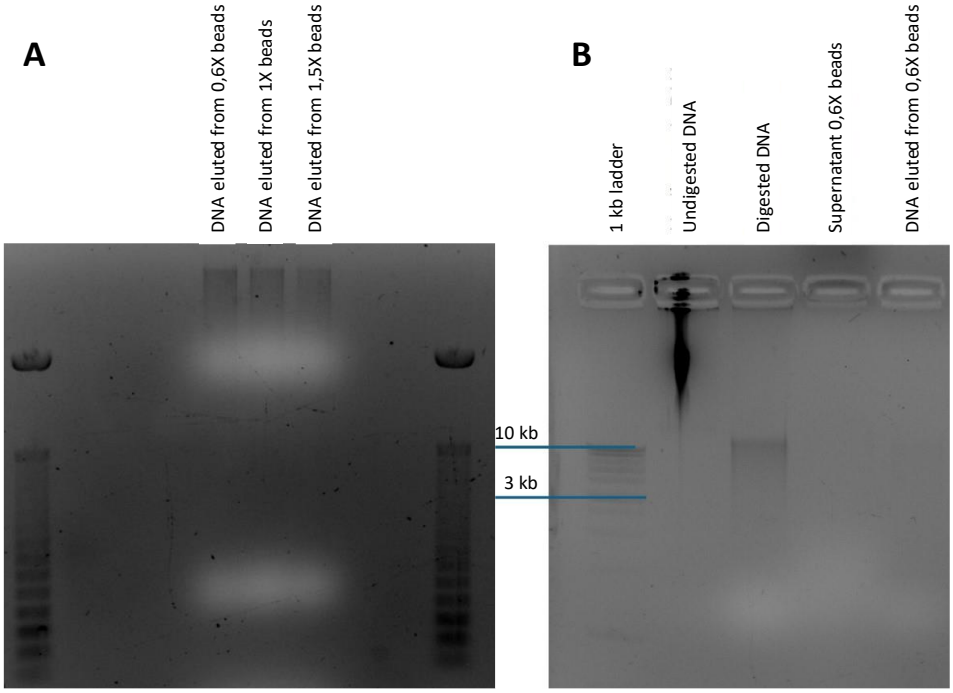

Figure 6. Gel images. A) The DNA eluted from different ratios of beads. B) Undigested DNA, digested DNA, the supernatant after the digested DNA was treated with 0,6X beads, and the DNA eluted from the 0,6X beads.

Sequencing results

After determining the best steps for the ddRAD protocol, the samples were prepped for Illumina sequencing. A total of 44 Gb paired-end reads were generated. Since the samples were pooled, approximately 22 Gb should have been generated per sample. In total 195,5 million reads were generated for the Kloosterbos sample, and 103,0 million reads for the Nijeveen sample. The reads were mapped to assembly 6 using Bowtie2 or BWA MEM (Table 4). The percentages of mapped reads were measured to assess the difference between the two mappers and to provide an indirect indication of the quality and completeness of the reference genome. The percentage of mapped reads with a map quality (MQ) above 0 is comparable across all samples and read mappers. The percentage of reads mapped with an MQ of 0 varies strongly among different read mappers, with BWA MEM showing a substantially higher percentage than Bowtie2. The opposite is observed for the percentage of unmapped reads, where the percentage of Bowtie2 is substantially higher compared to BWA.

Table 4. The distribution of unmapped reads, mapped reads with a map quality (MQ) higher than 0, and mapped reads with a MQ of 0.

Sample	**Kloosterbos**		**Nijeveen**	
Read Mappers	**BWA MEM**	**Bowtie2**	**BWA MEM**	**Bowtie2**
Mapped (with MQ>0)	65,1%	79,0%	68,1%	70,7%
MQ0	34,6%	5,3%	31,8%	6,2%
Unmapped	0,3%	15,7%	0,1%	23,2%

Variant calling

After obtaining the RAD sequencing data, the reads were mapped to assembly 6 (Table 3). To detect SNPs, five different approaches were tried using BCFtools or Stacks. A large variation in the number of detected SNPs was observed (Figure 7). Overall, BCFtools outperformed Stacks across all approaches. The number of SNPs detected by BCFtools was comparable when only one sample (Nijeveen) was used for variant calling and when two samples (Nijeveen and Kloosterbos) were used for variant calling. In

both cases, the detected SNPs corresponded to genuine variants when visually inspecting random SNPs in JBrowse. The Stacks ref_map.pl pipeline yielded only a very limited number of SNPs. Moreover, when inspecting random SNPs in JBrowse, they were not represented as true polymorphisms. This suggests that this approach produced largely false calls. The denovo_map.pl pipeline of Stacks generated more SNPs than the reference-based approach which was unexpected.

To detect true SNPs that can distinguish different wild daffodil populations, the detected markers in Nijeveen should preferably be homozygous to the reference for the Kloosterbos sequencing data. After filtering the SNPs called by BCFtools, a total of 1520 SNPs remained, which were homozygous alternative in Nijeveen, and homozygous to the reference for the Kloosterbos (Figure 7B). With the same filtering for the SNPs called by Stacks, zero SNPs remained.

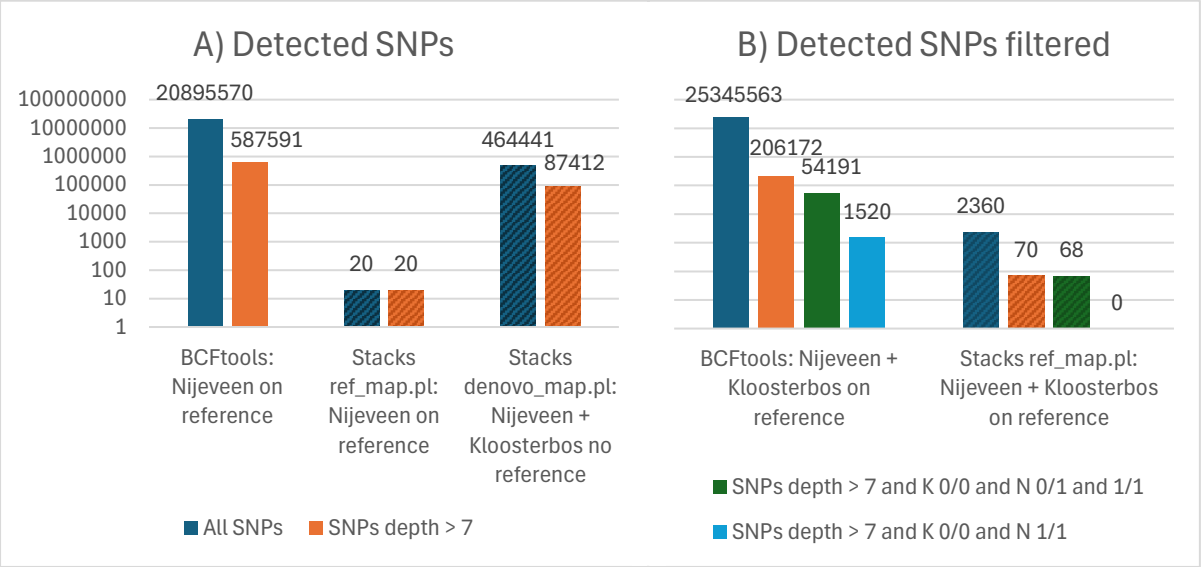

Figure 7. Number of detected SNPs across different calling approaches. A) Number of SNPs detected when only Nijeveen was mapped on the reference with BCFtools or Stacks using the ref_map.pl pipeline, or the number of SNPs detected using the denovo_map.pl pipeline from Stacks using no reference. B) Number of SNPs detected when RADseq data from Nijeveen and the Kloosterbos were both used as a sample and mapped on the reference genome. The filtering of the SNPs detected when the RADseq data from Nijeveen and the Kloosterbos was used was done with two different levels of stringency. Green shows the number of SNPs when the RADseq from the Kloosterbos was homozygous to the reference and the RADseq of Nijeveen was heterozygous or homozygous alternative. Light shows the number of SNPs when the RADseq from the Kloosterbos was homozygous to the reference and the RADseq of Nijeveen was homozygous alternative.

Discussion

During this research, a reference genome for the wild daffodil was assembled. Additionally, SNPs were detected that are possible markers by which different wild daffodil populations could be distinguished. This was done by first sequencing the whole genome of the wild daffodil sampled in the Kloosterbos. On this sequencing data, different analyses were performed to assess the quality (Figure 1) and to get an estimation of the genome size (Figure 2). With this information, different assemblies were made using the sequencing data (Table 3). The assembly with the best assembly statistics was chosen as the reference genome for the wild daffodil. This reference genome was screened for repetitive elements (Figure 4) and used for the variant calling. This variant calling was done with ddRAD sequencing data from the wild daffodil sampled in Nijeveen, to detect markers.

From the flow cytometry results (Table 1), it was clear that the genome from Tête-à-Tête was ~10% bigger than that of the wild daffodil from the Kloosterbos. It was expected that the genome from Tête-

à-Tête would be ~50% bigger than that of the wild daffodil growing in the Kloosterbos, since Tête-à-Tête is a known triploid and wild populations from *Narcissus* are often diploid (Pustahija *et al.*, 2024). Tête-à-Tête is an allotriploid containing 24 chromosomes (2n = 3x) (Sun *et al.*, 2024) with an average nuclear DNA content of 39,21 pg (Sochacki *et al.*, 2022). *Narcissus pseudonarcissus ssp. Pseudonarcissus L.* is a diploid containing 14 chromosomes (2n) with an average nuclear DNA content of 23,8 pg (Zonneveld, 2008). These measurements show that the genome of Tête-à-Tête is ~65% bigger than that of *N. pseudonarcissus ssp. Pseudonarcissus L.*. Although these nuclear DNA content measurements were conducted in different studies, the calculated 65% and the 10% measured in the results are far apart, suggesting that the wild daffodil from the Kloosterbos is not a *Narcissus pseudonarcissus ssp. Pseudonarcissus.* Assuming this, it could be possible that the wild daffodil from the Kloosterbos is not a diploid, but a triploid. However, this was not observed when looking through the alignment data from the RADseq of the Kloosterbos mapped on the reference genome. It could be that the wild daffodil from the Kloosterbos contains multiple B chromosomes. In *Narcissus poeticus,* 9 out of 13 populations contained 1 to 3 B chromosomes. It is known that Tête-à-Tête sometimes contains one B chromosome (Pustahija *et al.*, 2024). If the wild daffodil sampled in the Kloosterbos contained multiple B chromosomes, this could explain why the expected size difference was less than 50%. To confirm if the daffodil from the Kloosterbos is a diploid containing multiple B chromosomes, a karyogram can be made. Additionally, the nuclear DNA content can be measured and compared to that of other *Narcissus* species.

The first flow cell yielded 83,28 Gb and the second flow cell yielded 83,06 Gb (Table 2), which is slightly lower than the average yield of ~90 Gb reported by Goodwin and McCombie, (2019) for plant genome sequencing on a PromethION. This difference could be explained by different factors, including the variation in sequenced plant species, sample quality, or library preparation efficiency. Although the yield is below the average, it is within a comparable range. The N50 from the first flow was 2,6 kb higher than the N50 from the second flow cell. This slightly lower N50 may reflect the different library that was used after run 1 in the second flow cell, but overall, the N50 is comparable. The consistency in yield and N50 between the two flow cells suggests a good reproducibility of the library preparation and sequencing methodology.

A 39,6% increase was measured in average Phred read quality score when basecalling with the super accurate (SUP) model compared to the high accuracy (HAC) model. This can be partially explained by the fact that the average Phred quality score was calculated using SUP basecalled reads, from which reads with an average Phred read quality score below 10 were discarded (Zhao *et al.*, 2023). And the average Phred quality score, calculated with HAC basecalled reads, was determined using reads that were discarded due to an average Phred read quality score below 7. However, the filtering does not entirely explain the 39,6% increase in average Phred read quality score. The improvement can also be attributed to the use of the SUP model for basecalling. The improvement in Phred read quality score when comparing the HAC and SUP model was also observed in another study (Kuśmirek, 2023). To further improve the Phred read quality score, error correction using, for instance, HERRO, could have been done (Stanojevic *et al.*, 2024). The quality of the reads could also be improved by using error-correction algorithms that use high-accuracy short reads to correct the ONT long reads. For this, the ddRAD sequencing data from the Kloosterbos could be used (Wang, Zhao, *et al.*, 2021), although it is not known how successful this will be since the RADseq does not cover the entire genome.

From the k-mer histogram made by GenomeScope 2, it was calculated that the estimated genome size of the wild daffodil is 13,6 Gb (Figure 2B). The estimated genome size is smaller than the hypothesized approximate size of 15 Gb, but overall comparable. The full model of GenomeScope2 did not fit the observed histogram, and the error line did not resolve (Figure 2). This is also observed in other studies (Chen *et al.*, 2022; Ahuja *et al.*, 2024; Velotta *et al.*, 2025), and could have different explanations. Ranallo-Benavidez *et al.*, (2020) mentioned that the coverage must be at least 15X so it can resolve the error peak. From the observed peak, a coverage of 12 was calculated, which is too low and could explain why the error line was not resolved. It was also mentioned that GenomeScope only supports low error short read sequencing. The k-mer counting was done on ONT long read sequencing reads, which generally have a higher error rate, which could explain why the full model fit is poor. This also corresponds with the fact that a peak is present in the observed k-mer histogram of reads basecalled with the SUP model compared to the absent peak in the k-mer histogram of reads basecalled with the HAC model (Figure 2), since the reads the SUP model yielded a higher average Phred read quality score. Velotta *et al.*, (2025) suggest that the pore model fit of GenomeScope2 was due to high GA repeat content. This simple repeat content was measured at 5,12% of the assembly, which was a total of 46.246.753 base pairs. Our assembly contains 0,32% simple repeats (Figure 4), which is a total of 33.953.718 base pairs. The percentage of simple repeats of our assembly is lower, but the number of base pairs is in the same order. This suggests that the simple repeat content of our assembly may have played a role in the pore model fit of GenomeScope2. In the k-mer histogram of reads basecalled with the SUP model, only one peak can be observed (Figure 2B). The occurrence of one peak could be due to the wild daffodil sampled in the Kloosterbos being extremely homozygous. It is also possible that due to low coverage, the heterozygous and homozygous peaks are not resolved, making it impossible to determine if the observed peak is due to heterozygosity or homozygosity. When calculating the genome size from the heterozygous peak, only half of the genome is estimated (Hesse, 2023). If the observed peak is the heterozygous peak, this would mean that the estimated genome size of the wild daffodil is 27,2 Gb. However, it is unlikely that the genome size of the wild daffodil is 27,2 Gb when 79,7% complete BUSCO genes are present in an assembly of 10,8 Gb (Table 3). To achieve a more accurate estimation of the genome size of the wild daffodil, additional sequencing data should be generated. Additionally, to attempt a better model fit by GenomeScope, Hesse, (2023) recommended using the first version of GenomeScope when analyzing diploid species. Besides, it was recommended to set the maximum k-mer coverage value to fit the species instead of leaving it at default.

Assembly 6 was selected as the reference genome for the wild daffodil (Table 3). When comparing the assembly statistics of assembly 6 to the requirements of a high-quality genome assembly (Jung *et al.*, 2019; Lawniczak *et al.*, 2022; Wang and Wang, 2023), it can be concluded that assembly 6 is not of high quality. Although generating a high-quality assembly was not the primary aim of this research, it was a desirable side outcome. Assembly 6 can be classified as a poor/fair assembly when comparing the N50, the assembled genome coverage, and the percentage complete BUSCO genes (Jung *et al.*, 2019). This can be explained by different things, like a lack of coverage, the sequencing data containing too many errors, and the high content of repetitive elements (Figure 4) (Liao *et al.*, 2019). To tackle the question of whether a sufficient amount of data was generated, four graphs were made (Figure 3). It was observed that the 200% datapoint falls slightly below the linear trend observed from 60% of the dataset. To get a better estimation of whether the linear trend is leveling off, the graphs can be extended with information about 120%, 140%, 160% and 180% of the dataset.

Another factor could be that a different assembly algorithm or different parameter settings of Shasta would have been more fitting for this data. Shasta tends to produce a less contiguous and complete genome, which can be a result of over-splitting the genome due to repeats. The over-splitting by Shasta on our assemblies is very plausible, since the repeat content is very high and the assembly consists of a large number of contigs (146.889) (McCartney *et al.*, 2021; Wang, Chen, *et al.*, 2021; Espinosa *et al.*, 2024). To improve the assembly quality, different things can be done. As already mentioned, the Phred quality score of the reads can be improved by error correcting the reads. Assembly 6 could also improve with polishing using for instance, Pilon, Racon, or MEDAKA (McCartney *et al.*, 2021; Espinosa *et al.*, 2024). During this polishing step, the RAD sequencing reads from the Kloosterbos could be used, although it is not known how much this will improve the assembly since the RAD sequencing data does not cover the entire genome. To assess if there are contigs present in the assembly that are (partially) resembling each other, the assembly can be mapped back to itself. And to determine what the read depth is per contig and how the read depth is distributed over the contig, the ONT reads can be mapped back to the assembly. With this information, the best contigs can be selected that represent the wild daffodil.

Further noteworthy observations were that the Assemblies made by Shasta consistently outperformed the assemblies made by Miniasm (Table 3). Espinosa *et al.*, (2024) similarly reported that Shasta outperforms Miniasm in terms of completeness, based on comparisons of complete and fragmented BUSCO genes. However, the assemblies made by Miniasm were more contiguous compared to Shasta, which is contrasting when comparing our assemblies made by Miniasm and Shasta (Table 3). These differences may be attributed to variations in dataset characteristics, sequencing depth, or assembly parameters. The relatively strong performance of Shasta on our sequencing data may be attributed to its development for the human genome, which is ~3,2 Gb in size and contains a large proportion of repetitive elements (Shafin *et al.*, 2020). Miniasm, on the other hand, was tested on repeat sparse organisms and has not been optimized for large repeat-rich genomes (Li, 2016). This could explain the weaker performance of Miniasm on our data, particularly since assembly 6 consists of 91,97% repetitive elements (Figure 4).

Assembly 6 was used as a reference for the variant calling. Although it was stated earlier that this assembly is of poor/fair quality, it was hypothesized that genetic diversity could still be detected. This hypothesis is supported since SNPs were successfully identified in the Nijeveen sample. A complete reference genome would likely have reduced false positives and yielded additional confident SNPs (Aganezov *et al.*, 2022). Nonetheless, after filtering the detected SNPs, the outcome may still be confident, but with fewer SNPs when using a poor/fair quality reference genome. To improve accuracy, it is recommended to filter assembly 6 by excluding low-coverage contigs or regions and to perform purging to reconstruct a haplotype of the wild daffodil (Espinosa *et al.*, 2024). This will likely improve the mapping of the sequencing reads from other wild daffodil individuals, which will improve the variant calling.

Assembly 6 is largely composed of repetitive elements (91,97%), as hypothesized. The majority of the repetitive elements were classified as retro-elements (Figure 4). It should be noted that the repeat identification with RepeatModeler was carried out on assembly 5 (Table 3), rather than assembly 6. Additionally, the RepeatMasker analysis was conducted using the rush job option, which is ~10% less sensitive (Smit *et al.*). Therefore, the exact percentage of repeats may differ. Nonetheless, it is expected that the repeat content of assembly 6 falls within a similar range, since the repeat content

measured in garlic (91,3%) is very similar (Sun *et al.*, 2020). Garlic has a comparable genome size, is also diploid, and is part of the same order as the wild daffodil, which makes it plausible that, although very high, the wild daffodil has a repeat content higher than 90%. To identify the exact percentage of repetitive elements in assembly 6, RepeatModeler should be run using assembly 6 as the database, and during the RepeatMasker analysis, the slow search should be enabled.

The *in silico* digestion done on assembly 6 of the wild daffodil genome shows a clear shift towards smaller fragment sizes (Figure 5). These graphs provide a prediction of how the wild daffodil genome would be fragmented after digestion. However, because the *in silico* digestion was performed on a fragmented reference assembly, this prediction may not fully reflect the actual fragmentation pattern of the wild daffodil genome. Nonetheless, the *in silico* digestion still offers an approximation of the expected fragment sizes and their general distribution after digestion. Notably, there is an irregular pattern in the zoomed-in size distribution after digestion. The irregular pattern suggests that the digestion did not occur randomly, which likely reflects the positions of the restriction enzyme recognition sites. The high content of repetitive elements could explain this non-random distribution of the restriction enzyme recognition sites. Repetitive DNA often occurs in clusters or in uneven densities across the genome (Srivastava *et al.*, 2019). These repetitive regions may have fewer or more restriction enzyme recognition sites and could yield fragments that are highly similar in size.

The difference in eluted DNA among the tested bead ratios were minimal (Figure 6A). It was theoretically expected that a lower bead ratio would bind to the longest DNA fragments, resulting in shorter smears on the gel. However, no large difference in the length of the bands on the gel was observed, suggesting that the size-selection effect of the bead ratio may be less effective. Based on the slightly less elongated smear observed and how the beads theoretically should work, the 0,6X bead ratio was chosen to move forward with. DNA remaining in the supernatant after 0,6X bead treatment, presumed to contain mostly smaller fragments, was prepared for sequencing. Since the supernatant was below the detection limit of the gel (Figure 6B), the concentration was increased by treating the supernatant with 1,5X beads and eluting the DNA in a smaller volume.

The mapping results for the Kloosterbos and Nijeveen samples to the reference genome provide an insight into the success of the ddRAD sequencing and the alignments (Table 4). The proportion of reads mapped with a mapping quality (MQ) above 0 is relatively consistent across both samples and read mappers. This suggests that the library preparation was done successfully and consistently across samples. However, notable differences were observed between the two read mappers when comparing the reads mapped with an MQ of 0 and the unmapped reads. These large differences may be largely attributed to fundamental differences in how the two aligners calculate the mapping quality and how multi-mapping reads are handled (Giannoulatou *et al.*, 2014). The large fraction of unmapped reads of Bowtie2 suggests that the reference genome is not complete, which is in line with the 8,9% of missing BUSCO genes, and that the total length of the assembly does not correspond to the estimated length. This is in contrast to the small fraction of unmapped reads of BWA MEM, which suggests that the reference genome is fairly complete. Wu *et al.*, (2019) stated that BWA MEM is better at overall detecting more true-positive alignments and Bowtie2 is better at minimizing incorrect alignments. More research needs to be done on how the read mappers work on this data to understand the large differences in unmapped reads. Nonetheless, it is expected that the choice of read mapper does not significantly change the number of detected SNPs.

Five different strategies were carried out to detect SNPs that can distinguish the wild daffodil from Nijeveen from the wild daffodil from the Kloosterbos (Figure 7). The number of SNPs called with ref_map.pl pipeline from Stacks was very low compared to the SNPs called with BCFtools. It was also observed that the denovo_map.pl pipeline yields considerably more SNPs than the ref_map.pl pipeline from Stacks. This was unexpected, a comparable number of SNPs or more from the re_map.pl are normally observed in other studies (Shu and Moran, 2020; Bohling, 2020). It was hypothesized the ref_map.pl pipeline from Stacks could have a maximum read cut-off when it calls a SNP. However, this was not found in the literature, and other explanations were also absent. Overall, these findings suggest that the default options of the ref_map.pl pipeline from Stacks is not fitting for this data. It remains unclear what caused the difference in called SNPs and could be further investigated in future research.

BCFtools detected 1520 SNPs with a read depth higher than 7 (Figure 7). The ddRADseq data from the Kloosterbos sample needed to be identical at the positions where SNPs were identified in the Nijeveen ddRADseq data. This verification ensured that the remaining detected SNPs in the Nijeveen ddRADseq data did not result from errors in the assembly. This filtering step was therefore implemented to minimize the influence of potential technical errors and to ensure only true SNPs remain. This filtering step was not possible on the SNPs detected with only the Nijeveen sample mapped to the reference. The 1520 called SNPs could be good candidates as markers to detect genetic variation in other wild daffodils. It should be noted that the 1520 SNPs are homozygous in the Nijeveen sample. The 54.191 detected heterozygous SNPs in the Nijeveen sample are also possible markers for detecting genetic diversity. However, given that the read depth cut-off was 7, these SNPs are supported by only 3-4 reads per allele, which may reduce the reliability of these SNPs. Therefore, it is possible that alternative filtering strategies would have yielded additional SNPs that could also be useful for detecting genetic diversity. However, it is not known if the called SNPs are also a good marker to identify genetic variation in other individuals, as they were called with only one individual. Besides, the SNPs were not filtered on the quality, which could improve the markers and therefore detect genetic variability better.

It is currently uncertain whether the ddRAD sequencing protocol achieved the desired level of genome reduction. This should be determined to evaluate the suitability of the ddRAD sequencing protocol for future research. A cautious interpretation of the data suggest that the intended genome reduction may not have been fully reached, as visual inspection in JBrowse showed few genomic regions that were not covered by mapped ddRAD reads. However, only a fraction of the genome was examined, so this observation should be confirmed before drawing conclusions about the reached genome reduction. If the desired genome reduction was not reached the RADseq protocol should be revised, or other possibilities should be explored, like RNA sequencing. When the desired genome reduction is obtained, it is hypothesized that BCFtools could still identify reliable SNPs (Yao *et al.*, 2020), but other tools may also yield good results depending on the data that is used. The SNPs should be filtered differently with for instance a higher read depth cut-off, and the addition of also filtering for the variant quality score. After obtaining high-quality SNPs for multiple individuals, they can be processed in multiple ways to detect the genetic variability within and between wild daffodil populations.

Conclusion

During this research, a near-complete reference genome for the wild daffodil was assembled. Although the assembly is quite fragmented, it provides a valuable foundation for detecting the genetic variation within and between wild daffodil populations. The reference genome has already been applied during variant calling with ddRAD sequencing data of another wild daffodil individual, underlining that the current assembly is sufficient for detecting SNPs. However, because the reference genome has not been polished, and (partial) contigs with low read coverage were not filtered out, errors could be present in the assembly. These errors could lead to identifying false or uncertain SNPs. To minimize falsely detected SNPs, it is recommended to filter out low-coverage contigs. Additionally, reconstructing a haplotype of the reference genome could result in better mapping of the sequencing reads of other wild daffodil individuals.

Furthermore, ddRAD sequencing was used to identify SNPs in another wild daffodil individual. It remains to be assessed how much the wild daffodil genome was actually reduced. Determining the achieved reduction is essential for evaluating whether the ddRAD protocol was effective. A cautious expectation is that the desired level of reduction may not have been achieved with ddRAD. This assumption is based on inspection of the ddRAD sequencing data mapped to the reference in JBrowse, where most regions showed low-depth coverage. However, only a fraction of the mapped ddRAD sequencing data was inspected. So it is recommended to quantify the achieved genome reduction with ddRAD. When the desired genome reduction is achieved, BCFtools is a good tool for identifying SNPs, although it is recommended that future analyses apply other filtering approaches for improving the SNP quality.

Despite these limitations, this research represents an important first step towards uncovering the genetic diversity present in wild daffodil populations. The generated reference genome and initial SNP detection efforts together form a foundation for future research. Ultimately, these developments are a good start for the re-establishment of strong, genetically diverse, and self-sustaining wild daffodil populations in Drenthe and Overijssel.

Acknowledgements

I would like to thank my supervisors, Joost Keurentjes, René Boesten, Rens Holmer, and Frank Becker, for their guidance during my thesis. I am thankful for the weekly meetings where I received valuable feedback and suggestions, and I especially appreciated Joost's consistent presence and input during these meetings. I am grateful to Rens for his help and input on anything bioinformatics related, I really appreciated that he always took the time to discuss my questions or when I was unsure how to handle my data. René has been a great help in my writing process and provided me with good advice for the structuring of my report. Frank always assisted me with lab-related work and was very supportive in thinking along with every step of the experiments, for which I am thankful.

I would also like to thank Dirk Matthijs Meijberg from Landschapsbeheer Drenthe, who made it possible to sample the wild daffodil in the Kloosterbos.

I am also very grateful for the friends I made during my thesis, who made this period much more enjoyable. In particular, I would like to thank Lisa Nederpel. I'm very happy to have gained her as my

friend. We shared many laughs, having someone to sit next to, chat with, and occasionally vent to made this experience not only easier but also much more fun.

Supplementary data

All supplementary data can be found in eLabJournal.

References

Aganezov S, Yan SM, Soto DC, Kirsche M, Zarate S, Avdeyev P, *et al.* (2022). A complete reference genome improves analysis of human genetic variation. *Science* **376**: eabl3533.

Ahuja N, Cao X, Schultz DT, Picciani N, Lord A, Shao S, *et al.* (2024). Giants among Cnidaria: Large Nuclear Genomes and Rearranged Mitochondrial Genomes in Siphonophores (D Lavrov, Ed.). *Genome Biology and Evolution* **16**: evae048.

Barrett SCH, Harder LD (2005). The evolution of polymorphic sexual systems in daffodils (*Narcissus*). *New Phytologist* **165**: 45–53.

Bohling J (2020). Evaluating the effect of reference genome divergence on the analysis of empirical RADseq datasets. *Ecology and Evolution* **10**: 7585–7601.

Caldwell J, Wallace TJ (1955). Narcissus Pseudonarcissus L. *The Journal of Ecology* **43**: 331.

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011). *Stacks* : Building and Genotyping Loci *De Novo* From Short-Read Sequences. *G3 Genes|Genomes|Genetics* **1**: 171–182.

Chen S (2023). Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. *iMeta* **2**: e107.

Chen Y, Shah S, Dougan KE, Van Oppen MJH, Bhattacharya D, Chan CX (2022). Improved Cladocopium goreaui Genome Assembly Reveals Features of a Facultative Coral Symbiont and the Complex Evolutionary History of Dinoflagellate Genes. *Microorganisms* **10**: 1662.

Colling G, Hemmer P, Bonniot A, Hermant S, Matthies D (2010). Population genetic structure of wild daffodils (Narcissus pseudonarcissus L.) at different spatial scales. *Plant Syst Evol* **287**: 99–111.

Corlett RT (2020). Safeguarding our future by protecting biodiversity. *Plant Diversity* **42**: 221–228.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, *et al.* (2011). The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158.

Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, *et al.* (2021). Twelve years of SAMtools and BCFtools. *GigaScience* **10**: giab008.

De Coster W, Rademakers R (2023). NanoPack2: population-scale evaluation of long-read sequencing data (C Alkan, Ed.). *Bioinformatics* **39**: btad311.

Driguez P, Bougouffa S, Carty K, Putra A, Jabbari K, Reddy M, *et al.* (2021). LeafGo: Leaf to Genome, a quick workflow to produce high-quality de novo plant genomes using long-read sequencing technology. *Genome Biol* **22**: 256.

Espinosa E, Bautista R, Larrosa R, Plata O (2024). Advancements in long-read genome sequencing technologies and algorithms. *Genomics* **116**: 110842.

Ewels P, Magnusson M, Lundin S, Käller M (2016). MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**: 3047–3048.

Formenti G, Theissinger K, Fernandes C, Bista I, Bombarely A, Bleidorn C, *et al.* (2022). The era of reference genomes in conservation genomics. *Trends in Ecology & Evolution* **37**: 197–202.

Gaio D, Anantanawat K, To J, Liu M, Monahan L, Darling AE (2022). Hackflex: low-cost, high-throughput, Illumina Nextera Flex library construction. *Microbial Genomics* **8**.

Giannoulatou E, Park S-H, Humphreys DT, Ho JW (2014). Verification and validation of bioinformatics software without a gold standard: a case study of BWA and Bowtie. *BMC Bioinformatics* **15**: S15.

Goodwin S, McCombie WR (2019). Sequencing Complex Genomes with PromethION Technology in a Core Setting. *J Biomol Tech* **30**: S36–S37.

Hesse U (2023). K-Mer-Based Genome Size Estimation in Theory and Practice. In: Heitkam T, Garcia S (eds) *Plant Cytogenetics and Cytogenomics*, Methods in Molecular Biology. Springer US: New York, NY Vol 2672, pp 79–113.

Hochkirch A, Bilz M, Ferreira CC, Danielczak A, Allen D, Nieto A, *et al.* (2023). A multi-taxon analysis of European Red Lists reveals major threats to biodiversity (N Dahanukar, Ed.). *PLoS ONE* **18**: e0293083.

Hochkirch A, Samways MJ, Gerlach J, Böhm M, Williams P, Cardoso P, *et al.* (2021). A strategy for the next decade to address data deficiency in neglected biodiversity. *Conservation Biology* **35**: 502–509.

Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P (2019). Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes. *Trends in Plant Science* **24**: 700–724.

Kardos M, Armstrong E, Fitzpatrick S, Hauser S, Hedrick P, Miller J, *et al.* (2021). The crucial role of genome-wide genetic variation in conservation.

Klein Gotink E (2023). Determining possible hybridization in N. pseudonarcissus subsp. Pseudonarcissus in Drenthe. BSc Thesis, Wageningen University & Research.

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

Kuśmirek W (2023). Estimated Nucleotide Reconstruction Quality Symbols of Basecalling Tools for Oxford Nanopore Sequencing. *Sensors* **23**: 6787.

Langmead B, Salzberg SL (2012). Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**: 357–359.

Lawniczak MKN, Durbin R, Flicek P, Lindblad-Toh K, Wei X, Archibald JM, *et al.* (2022). Standards recommendations for the Earth BioGenome Project. *Proc Natl Acad Sci USA* **119**: e2115639118.

Li H (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.

Li H (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

Li H (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics* **32**: 2103–2110.

Liao X, Li M, Zou Y, Wu F, Yi-Pan, Wang J (2019). Current challenges and solutions of ***de novo*** assembly. *Quant Biol* **7**: 90–109.

Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM (2021). BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes (J Kelley, Ed.). *Molecular Biology and Evolution* **38**: 4647–4654.

Marçais G, Kingsford C (2011). A fast, lock-free approach for efficient parallel counting of occurrences of *k* -mers. *Bioinformatics* **27**: 764–770.

McCartney AM, Hilario E, Choi S, Guhlin J, Prebble JM, Houliston G, *et al.* (2021). An exploration of assembly strategies and quality metrics on the accuracy of the rewarewa (*Knightia excelsa*) genome. *Molecular Ecology Resources* **21**: 2125–2144.

van der Meijden R, Odé B, Groen K (C.) LG, Witte F (J.-P) M, Bal D (2000). *Bedreigde en kwetsbare vaatplanten in Nederland. Basisrapport met voorstel voor de RodeLijst*.

Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A (2018). Versatile genome assembly evaluation with QUAST-LG. *Bioinformatics* **34**: i142–i150.

Olkkonen E, Löytynoja A (2023). Analysis of population structure and genetic diversity in low-variance Saimaa ringed seals using low-coverage whole-genome sequence data. *STAR Protocols* **4**: 102567.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012). Double Digest RADseq: An Inexpensive Method for De Novo SNP Discovery and Genotyping in Model and Non-Model Species (L Orlando, Ed.). *PLoS ONE* **7**: e37135.

Pustahija F, Bašić N, Siljak-Yakovlev S (2024). Karyotype Variability in Wild Narcissus poeticus L. Populations from Different Environmental Conditions in the Dinaric Alps. *Plants* **13**: 208.

Ranallo-Benavidez TR, Jaron KS, Schatz MC (2020). GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat Commun* **11**: 1432.

Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, *et al.* (2020). Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38**: 1044–1053.

Shu M, Moran EV (2020). Testing pipelines for genome-wide SNP calling from Genotyping-By-Sequencing (GBS) data for Pinus ponderosa.

Smit AFA, Hubley R RepeatModeler Open-1.0.

Smit AFA, Hubley R, Green P RepeatMasker Open-4.0.

Sochacki D, Podwyszyńska M, Machlańska A, Dyki B (2022). Nuclear DNA Content, Selected Morphological and Anatomical Traits of Narcissus Cultivars and Breeding Clones. *Agronomy* **12**: 648.

Sparrius LB, Odé B, Beringen R (2014). *Basisrapport Rode Lijst Vaatplanten 2012 volgens Nederlandse en IUCN-criteria*. FLORON Nijmegen: Nijmegen.

Srivastava S, Avvaru AK, Sowpati DT, Mishra RK (2019). Patterns of microsatellite distribution across eukaryotic genomes. *BMC Genomics* **20**: 153.

Stanojevic D, Lin D, Nurk S, Florez De Sessions P, Sikic M (2024). Telomere-to-Telomere Phased Genome Assembly Using HERRO-Corrected Simplex Nanopore Reads.

Sun Y, Zeng J, Liu S, Zhou S (2024). FISH and GISH reveal genome composition of popular Narcissus cultivars and the possible ways of their origin. *Euphytica* **220**: 82.

Sun X, Zhu S, Li N, Cheng Y, Zhao J, Qiao X, *et al.* (2020). A Chromosome-Level Genome Assembly of Garlic (Allium sativum) Provides Insights into Genome Evolution and Allicin Biosynthesis. *Molecular Plant* **13**: 1328–1339.

Vaser R, Šikić M (2020). Raven: a de novo genome assembler for long reads.

Velotta JP, Iqbal AR, Glenn ES, Franckowiak RP, Formenti G, Mountcastle J, *et al.* (2025). A Complete Assembly and Annotation of the American Shad Genome Yields Insights into the Origins of Diadromy (B Fraser, Ed.). *Genome Biology and Evolution* **17**: evae276.

Wang J, Chen K, Ren Q, Zhang Y, Liu J, Wang G, *et al.* (2021). Systematic Comparison of the Performances of De Novo Genome Assemblers for Oxford Nanopore Technology Reads From Piroplasm. *Front Cell Infect Microbiol* **11**: 696669.

Wang P, Wang F (2023). A proposed metric set for evaluation of genome assembly quality. *Trends in Genetics* **39**: 175–186.

Wang Y, Zhao Y, Bollas A, Wang Y, Au KF (2021). Nanopore sequencing technology, bioinformatics and applications. *Nat Biotechnol* **39**: 1348–1365.

Wu X, Heffelfinger C, Zhao H, Dellaporta SL (2019). Benchmarking variant identification tools for plant diversity discovery. *BMC Genomics* **20**: 701.

Yao Z, You FM, N'Diaye A, Knox RE, McCartney C, Hiebert CW, *et al.* (2020). Evaluation of variant calling tools for large plant genome re-sequencing. *BMC Bioinformatics* **21**: 360.

Zerebecki RA, Hughes AR (2025). Environmental Stress and Resource Availability Affect the Maintenance of Genetic Variation in a Dominant Marsh Plant (*Spartina alterniflora*). *Molecular Ecology* **34**: e17628.

Zhao W, Zeng W, Pang B, Luo M, Peng Y, Xu J, *et al.* (2023). Oxford nanopore long-read sequencing enables the generation of complete bacterial and plasmid genomes without short-read sequencing. *Front Microbiol* **14**: 1179966.

Zonneveld BJM (2008). The systematic value of nuclear DNA content for all species of Narcissus L. (Amaryllidaceae). *Plant Syst Evol* **275**: 109–132.