Genetic variation of an apomictic dandelion lineage studied through *de novo* transcriptome assembly

M.Sc. Thesis
Nikolaos Pappas

January 2015
Genetic variation of an apomictic dandelion lineage studied through de novo transcriptome assembly

Nikolaos Pappas

January 2015

M.Sc. Thesis
Bioinformatics Group

WAGENINGEN UNIVERSITY

Supervisor: Prof. Dr. D. de Ridder
External Supervisors: Dr. J. Ferreira de Carvalho
                  Dr. K. Verhoeven
Abstract

Next-generation sequencing has enabled researchers to study non-model organisms in an unprecedented depth. Common dandelion (*Taraxacum officinale*) can play an important role in the study of apomixis and ecological epigenetics. Despite its importance, little genomic information is currently available. To this end, RNA-seq data from five accessions of the same dandelion apomictic lineage were *de novo* assembled using Trinity and Velvet/Oases. Based on several metrics, the resulting assemblies showed (i) high level of completeness in their biological content and (ii) discrepancies in the reported number of genes and transcripts. The effects of different parameter settings on the performance of these two assemblers are discussed in the context of *de novo* transcriptome assemblies of the budding yeast and the fission yeast. We then used a Trinity *de novo* assembly consisting of 123,322 transcripts distributed in 77,433 Trinity predicted genes, to call SNPs in 5 accessions of the same dandelion apomictic lineage. A total of 218,761 high quality SNPs were called, of which 120,864 are shared among all accessions. The subsequent clustering analysis based on the presence of the SNPs in the different accessions confirmed that the genetic variation between members of a clonal lineage is low. This report details a first attempt to gain insight in the genetic variation within an apomictic *T. officinale* lineage, through a *de novo* transcriptome assembly.
# Table of Contents

Abstract ............................................................................................................................................. i

1. Introduction .................................................................................................................................... 1

2. Materials and methods .................................................................................................................. 4
   2.1. RNA-seq Data ......................................................................................................................... 4
   2.2. Data preprocessing .................................................................................................................. 4
   2.3. De novo transcriptome assemblies .......................................................................................... 4
   2.4. Assemblies of Saccharomyces datasets .................................................................................. 5
   2.5. Assembly validation ............................................................................................................... 5
   2.6. Post-processing of the assemblies ......................................................................................... 6
   2.7. Variant Calling ....................................................................................................................... 6

3. Results and Discussion .................................................................................................................. 8
   3.1. De novo transcriptome assembly ............................................................................................ 8
       3.1.1. Assemblies vary substantially in number of transcripts and genes identified ................... 8
       3.1.2. Velvet/Oases produces an unrealistically high number of transcripts ......................... 9
       3.1.3. Redundancy reduction of the assemblies has an extreme effect ......................... 10
       3.1.4. Homology-based criteria are more informative for the assessment of transcriptome assemblies than length-based ones ................................. 10
       3.1.5. K-mer size and number have a prominent effect on the assembly size ......... 11
   3.2. Variant calling and sequence variation in a de novo transcriptome assembly ................. 16
       3.2.1. Mapping to the reference assembly .............................................................................. 16
       3.2.2. Base score recalibration increases confidence in SNP calls .................................. 17
       3.2.3. Filtering called SNPs reduces their number to a high-confidence subset ............. 18

4. Conclusions ..................................................................................................................................... 21
1. Introduction

Common dandelion, *Taraxacum officinale*, is a perennial weed widely distributed throughout Europe. It has been introduced by humans in all continents, except for Antarctica, and it is commonly found in meadows, disturbed grounds and woods (Van Dijk et al., 2003). Besides the sexual mode of reproduction typical for angiosperms, dandelion variants have been shown to reproduce asexually, via apomixis (Nogler, 1984). Apomixis refers to the clonal reproduction through seeds (Nogler, 1984) and in the dandelion it occurs through (i) the formation of an unreduced egg cell, after circumvention of normal meiosis during megasporogenesis (ii) the parthenogenetic embryo development and (iii) the autonomous endosperm formation (Van Dijk et al., 1999). Sexual variants are always diploid (2n=2x=16) and self-incompatible, while apomicts are always polyploid, commonly triploid (3x=24).

While mixed populations of both sexual and asexual *Taraxacum* are distributed in the southern part of the European continent, its northern regions are dominated by asexuals (Van Dijk et al., 2003). Several hypotheses have been proposed for the explanation of the success of the apomicts (Van Dijk et al., 2003). While many different apomictic genotypes can coexist within a population, the asexually reproducing apomicts produce clonal offspring, thus they are expected to show low genetic variation within lineages. However, evidence from morphological studies, as well as studies using different allozymes (variant forms of enzymes encoded by different alleles at the same locus) and DNA markers are strongly indicative of the contrary (van Baarlen et al., 2000). More recently, it was proposed that heritable epigenetic effects from the parent plant to the clonal lineage, such as DNA methylation, can also contribute to the wide range of phenotypic plasticity that is observed in the apomictic populations (Verhoeven et al., 2010a), complicating the efforts for better understanding the source of variation.

The co-occurrence of sexual and asexual dandelion populations, along with the observed genetic variation within the asexual lineages, render this species complex a suitable model for the study of the long-lasting evolutionary question of sexual and asexual reproduction - why does asexual reproduction arise in certain species, while sexual reproduction is more beneficial in the long-term (Meirmans et al., 2012)? In apomictic lineages neutral mutations in individual alleles accumulate faster, due to asexual reproduction, leading to increased heterozygosity in the short-term (Birky Jr, 1996). This higher than expected diversity within these lineages, in combination with their ability to start a new population from a single individual is considered an important factor for their success in colonising new areas (Paun et al., 2006). However, the asexual mode of reproduction can have a negative impact in the long-term adaptation of these organisms, due to lack of diversity and accumulation of mutations.
Despite the scarcity of the studies investigating the variability within clonal lineages, there is compelling evidence that suggest that genetic differentiation can arise through mutation accumulation (Mes et al., 2002). In principal, there should be two sources of variation present in such a lineage: (a) allelic variation that is shared among the different members and reflects changes in the mother plant; (b) genetic variation between members of the same lineage that resulted from mutations that occurred after their emergence. Traditional approaches to studying genetic variability between clonal members include the use of molecular markers such as the internal transcribed spacer (ITS) region, Amplified Fragment Length Polymorphism (AFLP) and microsatellite analysis (Mes et al., 2002; Verhoeven et al., 2010b).

Over the past years, Next Generation Sequencing (NGS) technologies have been successfully employed to generate valuable genomic resources for both model and non-model organisms (Mardis, 2008). Sequencing, assembling and annotating the full genomic sequence (DNA-seq) of a higher eukaryote, especially that of a polyploid plant, is a complex, time and money consuming effort. Some of the challenges during genome reconstruction are the sheer size of such genomes and the resolution of their highly repetitive content (Góngora-Castillo and Buell, 2013). Conversely, RNA-seq can provide useful information regarding the gene space and hence can be used as a fast and cheap alternative to whole genome sequencing for genotyping purposes (Haas et al., 2013a). In this context, a transcriptome assembly can serve as a valuable first step in the efforts of discovering genetic characteristics related with gene expression and studying variation through single nucleotide polymorphism (SNP) discovery (Ekblom and Galindo, 2011).

However, the reconstruction of the full transcriptome of an organism is not a trivial task to undertake. Major issues that need to be addressed include the differences in sequencing coverage between genes that are expressed at varying levels across tissues and conditions, and the presence of different transcript isoforms of the same gene due to alternative splicing (Góngora-Castillo and Buell, 2013).

In general, three strategies can be employed for transcriptome assembly, depending on the availability of a reference genome: a reference-based strategy, a de novo or a combination of the two, that all have their advantages and disadvantages (Martin and Wang, 2011).

Yet despite the fact that dandelion can serve as a model organism for studying the genetic variation within clonal lineages, little genomic resources are available for T. officinale in the public databases. As of yet, no reference genome is available, 597 nucleotide sequences are available in the NCBI repositories and gene expression data are available as 41,296 EST sequences. This effectively rules out the use of a reference-based assembly.
The aim of this study was to identify and quantify the genetic variation between members of a *T. officinale* clonal lineage. We therefore set out to produce a *de novo* transcriptome assembly for *T. officinale*, using RNA-seq data produced with the Illumina Hi-Seq platform. Using the assembly as a reference we further investigated the genetic variation between five accessions of an apomictic clonal lineage (section Macranthoides) of this organism through SNP detection. Based on the most reliable SNP calls, distances between the accessions were quantified in order to examine their relationships.
2. Materials and methods

2.1. RNA-seq Data

Seeds from dandelion individuals belonging to the same apomictic lineage (Macranthoides), confirmed via microsatellite analysis, were collected from 3 different geographic locations in central Europe (Germany, Czech Republic). After propagation for one generation in the greenhouse, seedlings were grown under two different conditions, in the greenhouse and in the field. Total RNA was extracted from leaf tissue. Four libraries were prepared for each accession per condition, resulting in a total of 40. Sequencing was carried out on an Illumina HiSeq platform using paired-end reads of length 101bp, with an average insert size of 275bp and resulted in the generation of 560,001,638 pairs of reads (Table 1). The sequencing data used in this study were kindly provided by J. Ferreira de Carvalho and K. Verhoeven, from KNAW-NIOO, the Netherlands Institute for Ecology.

Table 1. Number of reads that were generated and used in this study for the de novo assembly of the T. officinale transcriptome.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Field</th>
<th>Greenhouse</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>64,096,436</td>
<td>47,506,100</td>
<td>111,602,536</td>
</tr>
<tr>
<td>8</td>
<td>49,282,921</td>
<td>45,846,256</td>
<td>95,129,177</td>
</tr>
<tr>
<td>11</td>
<td>67,198,024</td>
<td>47,822,541</td>
<td>115,020,565</td>
</tr>
<tr>
<td>12</td>
<td>70,713,183</td>
<td>46,447,026</td>
<td>117,160,209</td>
</tr>
<tr>
<td>13</td>
<td>69,802,225</td>
<td>51,286,926</td>
<td>121,089,151</td>
</tr>
</tbody>
</table>

560,001,638

2.2. Data preprocessing

Raw reads from the 40 samples were pre-processed with the FASTQC software (v. 0.10.1) using default settings (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). The first 10 bases from each read were trimmed off using the FASTQ Trimmer utility available with the FASTX toolkit (v. 0.0.13.1, parameters "-Q33 -f 11") (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The resulting 91bp long reads from all the samples were combined in two FASTQ files, each containing one member of the read pairs.

2.3. De novo transcriptome assemblies

Two de novo transcriptome assemblers were used: Trinity r2014-04-13p1 (Grabherr et al., 2011) and Velvet v1.2.10 /Oases v0.2.08 (Zerbino and Birney, 2008; Schulz et al., 2012). The trimmed reads were also normalized in silico using the Trinity normalization utility before assembly at a k-mer coverage of 30 (option --
normalize_by_max_read_cov). This process eliminates reads from the read set that have a median k-mer abundance larger than the predefined threshold and can significantly reduce the size of the dataset. This reduction can in turn help reduce run times and memory requirements, without affecting the quality of the final assembly (Brown et al., 2012).

Both the normalized and unnormalized read sets were assembled with Trinity at a k-mer size of 25. For the unnormalized reads the –iworm_kmer_cov option was set to 2, in order to eliminate possible erroneous k-mers. Velvet/Oases supports assemblies at different k-mer lengths, so it was used to assemble k-mers from 21 to 91 with a step size of 10. It was not possible to run Velvet/Oases on the unnormalized reads set due to memory limitations. Contigs smaller than 200bp were discarded from all assemblies.

2.4. Assemblies of Saccharomyces datasets

In order to assess the effect of different parameters on the number of assembled transcripts, de novo transcriptome assemblies were performed for two well characterized eukaryotic model organisms, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Previously published RNA-seq data for *S. cerevisiae* were obtained from (Nookaew et al., 2012) and for *S. pombe* the dataset provided by (Haas et al., 2013) was used.

First, the parameters of interest for Trinity were identified. For the first phase of contig assembly with the Inchworm algorithm, the minimum k-mer coverage parameter (--min_kmer_cov, default value 1) and for the third step of de Bruijn graph simplification with the Butterfly algorithm, the minimum overlap of reads with the growing transcript path (--path_reinforcement_distance, default value 75 for paired end reads), the minimum identity percentage for two paths to be merged into a single path (--min_per_id_same_path, default value 95), the maximum allowed differences encountered between path sequences to combine them (--max_diffs_same_path, default value 2) and the maximum number of internal consecutive gap characters allowed for paths to be merged into single paths (--max_internal_gap_same_path, default value 10).

For the Velvet/Oases assembler, the effect of using different k-mer sizes from 17 to 25 with a step size of 2 was assessed.

2.5. Assembly validation

The quality of the assemblies was assessed based on the following metrics:

- Read mappability: reads were mapped to the assemblies using Bowtie2 v2.1.0 (Langmead and Salzberg, 2012) with default settings. The percentage of mapped reads was used as the metric.
• Homology: standalone NCBI BLAST v2.2.25+ (Altschul et al., 1997) was used to perform BLASTx comparisons against the UniProtKB A. thaliana proteome (April 2014) set and BLASTn against a reference EST set, available from http://compgenomics.ucdavis.edu/compositae_overview.php. BLAST searches were performed with an E-value cutoff at 1e-5. The number of transcripts with a significant hit was used as the metric.

• Core gene set retrieval: a CEGMA analysis (Parra et al., 2007) was also carried out in order to detect the proportion of a highly conserved core eukaryotic genes found back in the assemblies (Parra et al., 2009). In absence of experimental data about the state of the gene space, the CEGMA analysis tool can be used to assess the biological content of the assembly. While originally developed for draft genome assemblies, it can be used on transcriptome assemblies. The percentage of partial and complete proteins retrieved was used as the metric.

2.6. Post-processing of the assemblies

In order to reduce redundancy, the raw assemblies were first processed with CD-HIT-EST (Li and Godzik, 2006) to remove any identical transcripts (option –c 1) and then with the TGICL tool (Pertea et al., 2003). The TIGR Gene Indices clustering tool was originally developed for the analysis of Expressed Sequence Tags (ESTs). The sequences are first clustered based on their pair-wise sequence similarity and then assembled per cluster to produce longer, more complete consensus sequences. The per cluster sequence assembly is carried out by CAP3, a DNA sequence assembly program (Huang and Madan, 1999). Two sets of parameters were used, default and “-p 99 -l 50 -v 100” where -p is the minimum percent identity for overlaps (default 94), -l is the minimum overlap length (default 40) and -v is the maximum length of unmatched overhangs (default 30) (Nakasugi et al., 2014).

For subsequent analyses, a Trinity de novo transcriptome assembly produced by J. Ferreira de Carvalho et al. (in preparation) was used as a reference, consisting of 123,322 transcripts distributed in 77,433 Trinity ‘genes’, in order to facilitate inferences and comparisons with the results of their ongoing study.

2.7. Variant Calling

For SNP calling, we used the GATK workflow (McKenna et al., 2010). The individual read files with the first 10 bases trimmed from all samples and both conditions were merged per accession. The resulting FASTQ files were mapped to the reference Trinity transcriptome assembly using the BWA-backtrack program (Li and Durbin, 2009). A gap opening penalty was set to 6 to allow for more leniency during mapping (parameter –o 6). Mapping statistics were calculated with the samtools flagstat utility (Li et al., 2009).
The resulting alignments were further preprocessed using Picard tools (http://broadinstitute.github.io/picard). In particular, read group information was added for every accession, duplicates were marked and the final BAM files were sorted and indexed.

In order to call SNPs using the GATK analysis pipeline (McKenna et al., 2010), a part of the pre-processing requires the recalibration of sequence quality scores in aligned BAM files, making use of a pre-existing database of known polymorphisms. In absence of this database a new one can be bootstrapped by calling SNPs with the unrecalibrated data at first. From this original SNP database only the ones with the highest confidence are kept. For our dataset we used a quality threshold of 100 (Nicolaï et al., 2012), as reported in the QUAL field of the resulting Variant Calling File (vcf). The SNPs present in this de novo created database are treated as known polymorphisms and are used to recalibrate the data. A new round of SNP calling is carried out with the recalibrated data and the whole process is repeated until convergence. In our case, this was done for 0.1% of the reads in the original BAM files, since it has been shown that downsampling can improve significantly the recalibration speed, while keeping the relative information (http://gatkforums.broadinstitute.org/discussion/44/base-quality-score-recalibration-bqsr). The resulting VCF files for each iteration, were filtered based on the presence of only single nucleotide polymorphisms with a value in the QUAL field larger than 100. This process was repeated 3 times for each accession and the final recalibration parameters obtained from the downsampled BAM files were used to recalibrate the original raw BAM files.

The recalibrated BAM file for each accession was then used for the final step of SNP calling, using Haplotype Caller. Hard filters were applied on the resulting VCF files, according to the recommendations of the GATK group for variant calling from RNA-seq data, to ensure that the called variants were of the highest confidence possible. Specifically, only single nucleotide polymorphisms were kept both in the reference and the alternative fields, as reported in the vcf files, with a quality larger than 100. A table of all shared and unique SNPs for all 5 accessions was produced using a custom Python script, that was then analysed with RStudio (R version 3.0.1) to calculate a matrix of Euclidean and Jaccard distances between the 5 accessions. Based on the calculated distance matrices, the accessions were hierarchically clustered.
3. Results and Discussion

3.1. De novo transcriptome assembly

A variety of algorithms is currently available for de novo assembly of the transcriptome of non-model organisms that lack a reference genome. A number of considerations have to be taken into account, both prior to the assembly (data pre-processing) and during their application to the specific dataset. In this study we wanted to first get an assembly as complete as possible and then quantify the genetic variation between members of a dandelion apomictic lineage, in the form of SNPs. To this end, we first wanted to investigate how some of the parameters of the various algorithms affect the final result when they are changed in two of the most commonly used de novo transcriptome assemblers, Trinity and Velvet/Oases.

3.1.1. Assemblies vary substantially in number of transcripts and genes identified

Three different de novo transcriptome assemblies of T. officinale were produced using raw and normalized reads sets as input for Trinity and Velvet/Oases. An overview of the resulting assemblies and the metrics used for their validation is shown in Table 2.

After in silico normalization, the original number of 560 million pair reads was significantly reduced to a total of 16,541,221 paired reads. Using both the original and the normalized reads we were able to produce two Trinity assemblies. The resulting assembly using all the reads is herein referred to as TrAll and the one using the normalized reads as TrNorm. Surprisingly, even though the dataset was reduced to less than 3% of the original dataset, the number of assembled transcripts increased. A total of 44,569 more transcripts were assembled compared to the TrAll assembly and the number of predicted genes (Trinity components) increased from 77,763 to 84,855. This increase in the reported number of transcripts between assemblies using normalized and unnormalized reads has also been previously described in de novo transcriptome assemblies of mouse and yeast (Brown et al., 2012) and sea squirts (Lowe et al., 2014) and has been attributed to differences in the number of reported splice variants.
Table 2. Quality metrics and statistics of three de novo assemblies of the *T. officinale* transcriptome.

<table>
<thead>
<tr>
<th></th>
<th>TrAll</th>
<th>TrNorm</th>
<th>VelvOasNorm</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of transcripts</td>
<td>124,568</td>
<td>169,227</td>
<td>621,100</td>
</tr>
<tr>
<td>No. of “genes”/loci</td>
<td>77,763</td>
<td>84,855</td>
<td>34,538</td>
</tr>
<tr>
<td>N50</td>
<td>1,482</td>
<td>1,570</td>
<td>1,682</td>
</tr>
<tr>
<td>Mean transcript length</td>
<td>919.1</td>
<td>1,005.91</td>
<td>1,199.12</td>
</tr>
<tr>
<td>% GC</td>
<td>39.37</td>
<td>39.27</td>
<td>39.54</td>
</tr>
<tr>
<td>% complete CEGMA proteins</td>
<td>97.98</td>
<td>94.76</td>
<td>99.60</td>
</tr>
<tr>
<td>% partial CEGMA proteins</td>
<td>99.6</td>
<td>99.6</td>
<td>100.00</td>
</tr>
<tr>
<td>Mapped % of all reads (Bowtie2)</td>
<td>94.92</td>
<td>94.74</td>
<td>97.08</td>
</tr>
<tr>
<td># BLAST hits against <em>T. officinale</em> ESTs (%)</td>
<td>27,995 (22.5)</td>
<td>48,509 (28.7)</td>
<td>247,500 (39.8)</td>
</tr>
<tr>
<td># BLAST hits against <em>A. thaliana</em> proteome (%)</td>
<td>52,684 (42.3)</td>
<td>85,287 (50.4)</td>
<td>196,622 (31.6)</td>
</tr>
</tbody>
</table>

3.1.2. Velvet/Oases produces an unrealistically high number of transcripts

The Velvet/Oases assembly was produced using the normalized reads (hence VelvOasNorm) and yielded a very large number of transcripts (621,100), distributed across 34,548 loci. Ideally, all the transcripts in one locus correspond to different isoforms of the same gene (Schulz et al., 2012). The number of reported loci is close to the average number of genes that are expected to be found in a higher plant genome (Sterck et al., 2007), which prompted us to examine more in detail the high number of reported transcripts. We found that almost half of the assembled transcripts (361,280) were assigned to locus 1. Using CD-HIT-EST on these transcripts to further cluster together identical sequences reduced the number of transcripts to 278,565. A CEGMA analysis revealed that approximately 60% of the 248 core eukaryotic genes are represented in these sequences, while a BLAST comparison against the *A. thaliana* proteome resulted in 90,747 locus 1 contigs having a significant hit. According to Posnien et al. (Posnien et al., 2014), who encountered a similar discrepancy in their *de novo* transcriptome assembly using Velvet/Oases, this is a computational artifact, possibly due to sequencing errors or even naturally occurring polymorphisms. These can lead to small differences in the sequence of the assembled transcripts, which are treated as distinct by the assembler, but are similar enough to get clustered together in the same locus.
These results indicate that the process by which the Velvet/Oases assembler groups different transcripts in loci is very sensitive to sequence variation. From a biological standpoint this could hint to the presence of many different variants. Since the apomictic dandelions are triploid, there can be several different alleles in one locus which can accumulate different mutations over time. However, despite the fact that reporting half of the transcriptome in one locus is unrealistic, the results from the CEGMA and BLAST analyses suggest that transcripts similar to those in locus 1 are also present in other loci.

3.1.3. Redundancy reduction of the assemblies has an extreme effect

Next, we wondered how redundant the assemblies were, i.e. how many transcripts and genes are highly similar. We therefore used the TGICL tool which first clusters assembled sequences based on their pair-wise similarity and then try to assemble these sequences into longer, more complete contigs using CAP3 (Pertea et al., 2003; Huang and Madan, 1999). It has been previously successfully applied in transcriptome assemblies of the allopolypl oid plant *Nicotiana benthamiana* (Nakasugi et al., 2014). In our hands, this "cluster and assemble" strategy for redundancy reduction of the assemblies, reduced their size of the assemblies enormously. For example, when applied with default parameters, the number of transcripts in the TrNorm assembly was reduced from 169,227 to 6,348. Another attempt using a more stringent set of parameters to avoid dissimilar sequences from clustering together, still resulted in a reduction of the number of sequences to 10,367.

3.1.4. Homology-based criteria are more informative for the assessment of transcriptome assemblies than length-based ones

Other commonly used quality metrics for *de novo* assemblies, such as the N50 size (the size of a contig such that 50% of the assembled bases are included in contigs of this length or longer), mean transcript length and assembly size were similar for both the normalized and unnormalized Trinity assemblies. However, these metrics have been more commonly used for the assessment of genome assemblies and are less informative in the context of the transcriptome (O'Neil and Emrich, 2013). Instead of these length based criteria, others based on the content of the transcriptome assembly are more valuable in the effort of assessing its accuracy and completeness.

In this study, we used homology-based searches against a well curated database of *A. thaliana* protein sequences and a publicly available EST database of *T. officinale*, as well as CEGMA analyses. BLAST searches against the *A. thaliana* proteome resulted in a modest number of transcripts with a significant hit. For the TrRaw and TrNorm assemblies 52,684 (42.3%) and 85,287 (50.4%) transcripts, respectively, showed significant similarity to *A. thaliana* proteins. The remainder of the assembled
transcripts could correspond to non-coding RNAs or be artifacts caused by sequencing errors. The CEGMA results reveal that all three assemblies were able to capture almost all (TrAll and TrNorm) or all (VelvOasNorm) transcripts that correspond to genes conserved across all eukaryotes, affirming that there is valid biological information included in them. This shows that the assemblers, when applied to our dataset, exhibit high sensitivity since we were able to retrieve all of the Core Eukaryotic Genes. However, the high representation of sequences that are more likely erroneous, demonstrates that they have low specificity.

In summary, we were able to produce three different assemblies that contain most of the biological information, as was confirmed from their homology-based assessment. Even though we attempted to minimize the large variation in number of reported transcripts and genes, there is no clear evidence as to which assembly would be most suitable for further analyses. We therefore set out to better understand the effect of key assembler parameter settings on these results.

3.1.5. K-mer size and number have a prominent effect on the assembly size

Assemblers

The algorithms we used for assembly, Trinity and Velvet/Oases, are both based on similar principles but work in different ways, and parameters therefore will have different effects.

Trinity is a de Bruijn graph based assembler that reconstructs full length transcript sequences from short NGS reads in a three step process (Fig. 1). First, the input reads are partitioned into k-mers (short sequences of length k), which are then assembled into transcript sequences, based on the frequencies of the k-mers and a k-1 overlap. This process is carried out by an algorithm called Inchworm. Second, related contigs, in terms of sequence content, are clustered together and a de Bruijn graph is created for each cluster by a module called Chrysalis. Third, the Butterfly module makes use of information on reads that support the paths in the constructed graphs to resolve the alternative spliced isoforms and the plausible transcripts. (for a more detailed explanation of how Trinity works see (Grabherr et al., 2011).
Figure 1. Schematic representation of the three step process used by Trinity for transcriptome reconstruction. (a) From the initial read set (top) Inchworm constructs linear sequences (bottom) based on a k-mer graph (middle). (b) Chrysalis uses k-1 sequence overlaps between the Inchworm assembled sequences, pools them together and builds de Bruijn graphs for each pool. (c) Butterfly then processes the de Bruijn graphs (trimming and compacting) and makes use of the initial read information (dashed color lines) to report the final transcripts. Taken from Grabherr et al., 2011.

The Oases assembly process is quite similar to that of Trinity. The input reads are partitioned into k-mers and assembled into contigs by the Velvet genomic assembler. These contigs are then processed by Oases for error correction and their resolution into distinct transcript assemblies (Schulz et al., 2012).

An important difference between the two assemblers is the k-mer size used for the constructing of the de Bruijn graphs. Trinity has a single, fixed k-mer length (k=25) while Oases is able to produce multiple assemblies at varying k-mer lengths, that are then merged into one single final assembly. The use of multiple k-mer values in transcriptome assemblies is considered to be more advantageous, since it allows to better handle the varying expression levels of transcripts in RNA-seq data (Zhao et al., 2011). Most de Bruijn graph assemblers were developed for genome assembly and work with the assumption of even sequencing coverage. For these assemblers, the challenge has been the choice of a single, optimal k-mer length. However, for transcript assemblies the differences in coverage of the sequenced transcripts can
cause problems in their reconstruction and by varying the value of this critical parameter, a more complete assembly can be produced (Robertson et al., 2010). Transcripts with low abundance are represented by less reads in the RNA-seq data and can be more accurately reconstructed by using smaller k-mer lengths (Schulz et al., 2012).

**Parameter settings**

The parameters we examined in the Trinity assemblies, for impact on the reported numbers of transcripts are summarized in Table 3.

**Table 3.** Parameters examined for their effect on the number of reported transcripts in *de novo* transcriptome assemblies.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Description (Trinity module)</th>
<th>Default value</th>
<th>Alternative value tested</th>
</tr>
</thead>
<tbody>
<tr>
<td>min_kmer_cov</td>
<td>The minimum abundance of a k-mer used in the initial contig reconstruction (Inchworm)</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>path_reinforcement_distance</td>
<td>Minimum overlap of reads with growing transcript (Butterfly)</td>
<td>75</td>
<td>1</td>
</tr>
<tr>
<td>min_per_id_same_path</td>
<td>Minimum percentage of identity for two paths to be merged into a single path (Butterfly)</td>
<td>95</td>
<td>90</td>
</tr>
<tr>
<td>max_diffs_same_path</td>
<td>Maximum allowed number of differences encountered between path sequences to combine them (Butterfly)</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>max_internal_gap_same_path</td>
<td>Maximum number of internal consecutive gap characters allowed for paths to be merged into single paths (Butterfly)</td>
<td>10</td>
<td>15</td>
</tr>
</tbody>
</table>

*a Abbreviations and descriptions of the parameters are taken from the Trinity usage manual.

*b These parameters were changed simultaneously to produce one assembly.

**Yeast transcriptome assemblies**

In an effort to better understand the influence of different parameters on the assemblies produced by different algorithms, we turned to organisms for which a ground truth, in the form of a well-annotated genome, was available. We found transcriptomics data of two well characterized yeast strains, the budding yeast (*Saccharomyces cerevisiae*) and the fission yeast (*Schizosacharomyces pombe*). *S. cerevisiae* has a genome size of ~12 Mbp and is comprised of ~6,500 genes, distributed in 16 nuclear chromosomes (http://www.yeastgenome.org/)
genomesnapshot). The S. pombe genome is of similar size (12.5 Mbp) distributed in 3 nuclear chromosomes and has a smaller number of ~5,000 protein coding genes (http://www.pombase.org/status/statistics). Only 5% of the genes in S. cerevisiae contain an intron and virtually no alternative splicing is known to take place, with very few exceptions (Juneau et al., 2009), in contrast with the S. pombe where alternative splicing is more frequent (Wood et al., 2002). With this information, assembling the transcriptome of S. cerevisiae should yield a number of transcripts very similar to the number of genes, while the S. pombe should result in a higher number of transcripts, since alternate isoforms of the genes are expected to be present.

In total, 4 different assemblies were produced for each yeast dataset (Table 4). Each time, only one of the parameters was changed, with the exception of min_per_id_same_path, max_diffs_same_path and max_internal_gap_same_path which were changed simultaneously. The assemblies produced are referred to as Default (all parameters set to default values), Iworm5 (min_kmer_cov set to 5), PD1 (path_reinforcement_distance set to 1) and B90_10_15 (min_per_id_same_path set to 90, max_diffs_same_path set to 10, max_internal_gap_same_path set to 15).

**Table 4.** Basic statistics for de novo transcriptome assemblies produced for the two yeast datasets using Trinity.

<table>
<thead>
<tr>
<th></th>
<th>Default</th>
<th>Iworm5</th>
<th>PD1</th>
<th>B90_10_15</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S. cerevisiae</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># genes</td>
<td>9,817</td>
<td>4,860</td>
<td>9,803</td>
<td>9,821</td>
</tr>
<tr>
<td># transcripts</td>
<td>10,074</td>
<td>5,082</td>
<td>10,064</td>
<td>10,013</td>
</tr>
<tr>
<td>% GC</td>
<td>40.31</td>
<td>38.5</td>
<td>40.3</td>
<td>40.3</td>
</tr>
<tr>
<td>N50</td>
<td>3,828</td>
<td>3,105</td>
<td>3,858</td>
<td>3,822</td>
</tr>
<tr>
<td>Average contig length</td>
<td>1,394.68</td>
<td>2,153.54</td>
<td>1,397.45</td>
<td>1,388.6</td>
</tr>
<tr>
<td>Total assembled bases</td>
<td>14,050,008</td>
<td>10,944,293</td>
<td>14,063,969</td>
<td>13,904,077</td>
</tr>
<tr>
<td><strong>S. pombe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td># genes</td>
<td>8,786</td>
<td>6,454</td>
<td>8,783</td>
<td>8,785</td>
</tr>
<tr>
<td># transcripts</td>
<td>9,243</td>
<td>6,570</td>
<td>9,249</td>
<td>9,198</td>
</tr>
<tr>
<td>% GC</td>
<td>38.04</td>
<td>39.34</td>
<td>38.04</td>
<td>38.03</td>
</tr>
<tr>
<td>N50</td>
<td>1,592</td>
<td>1,097</td>
<td>1,588</td>
<td>1,593</td>
</tr>
<tr>
<td>Average contig length</td>
<td>1013.46</td>
<td>732.87</td>
<td>1,014.2</td>
<td>1,013.43</td>
</tr>
<tr>
<td>Total assembled bases</td>
<td>9,367,448</td>
<td>4,814,962</td>
<td>9,379,642</td>
<td>9,312,546</td>
</tr>
</tbody>
</table>

In general, the results are consistent with our observations on the dandelion assemblies. In all of the produced assemblies, except for the Iworm5, the number of genes and transcripts are largely overestimated. However, the number of transcripts per gene seems to be at acceptable levels, given that S. pombe shows a larger rate of alternative splicing than S. cerevisiae. It is noteworthy that for the S. cerevisiae dataset the total number of assembled bases exceeds the size of the genome.
Our results show that the parameter with the most prominent effect on the number of reported transcripts, for both the fission and budding yeast datasets, is the minimum abundance of the k-mers to be used in the assembly. Indeed, changing this parameter to a higher value is a recommended way of eliminating error containing k-mers. These k-mers are more abundant in the lower counts and can be removed without affecting the sensitivity of the transcripts reconstruction (http://trinityrnaseq.sourceforge.net/advanced_trinity_guide.html).

The PD1 and B90_10_15 assemblies exhibited few differences in their metrics, compared to the default assembly. Setting the path reinforcement distance parameter to 1 should result in a very lenient extension of the transcript, since only an overlap of 1 nucleotide in the reads that support that particular path is required for extension. We would then expect more paths to be supported during graph traversal and chosen for extension, even if they are erroneous, resulting in a higher number of reported transcripts. Conversely, by setting the three parameters for the BD90_10_15 assembly to values that should require more stringent path extension we would expect a decrease in the reported number of transcripts. However, this was not the case and the resulting assembly was quite similar to the Default.

We also produced 6 assemblies with varying k-mer lengths using the Velvet/Oases assembler (Table 5). The number of transcripts is again overestimated for all the assemblies, at levels that are even greater than those of Trinity. The number of predicted loci however seems to be underestimated. Even for the Merged assembly, which is meant to contain a non-redundant representation of all the single k-mer assemblies (Schulz et al., 2012), all reported statistics are overestimated.

Table 5. Basic statistics for de novo transcriptome assemblies produced for the budding yeast dataset using Velvet/Oases at different k-mer lengths.

<table>
<thead>
<tr>
<th></th>
<th>k17</th>
<th>k19</th>
<th>k21</th>
<th>k23</th>
<th>k25</th>
<th>Merged</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong># transcripts</strong></td>
<td>33,871</td>
<td>28,202</td>
<td>20,650</td>
<td>16,436</td>
<td>14,278</td>
<td>36,742</td>
</tr>
<tr>
<td><strong># loci</strong></td>
<td>3,129</td>
<td>3,657</td>
<td>4,011</td>
<td>4,162</td>
<td>4,000</td>
<td>4,177</td>
</tr>
<tr>
<td><strong>GC %</strong></td>
<td>39.36</td>
<td>38.7</td>
<td>38.7</td>
<td>38.72</td>
<td>38.7</td>
<td>38.59</td>
</tr>
<tr>
<td><strong>N50</strong></td>
<td>376</td>
<td>2,194</td>
<td>3,452</td>
<td>3,977</td>
<td>4,102</td>
<td>3,866</td>
</tr>
<tr>
<td><strong>Average contig length</strong></td>
<td>391.17</td>
<td>1098.1</td>
<td>1,753.54</td>
<td>2,247.06</td>
<td>2,257.98</td>
<td>2,324.31</td>
</tr>
<tr>
<td><strong>Total assembled bases</strong></td>
<td>13,249,381</td>
<td>30,968,717</td>
<td>36,210,515</td>
<td>36,932,654</td>
<td>33,255,480</td>
<td>85,399,753</td>
</tr>
</tbody>
</table>

In summary, the results of the de novo transcriptome assemblies from well characterized organisms seem to confirm the variability in performance of the two assemblers. We tried out different parameter settings for each of the algorithms, but we were not able to reach a consensus as to which of these are the best for our data. A possible suggestion would be to use Trinity with an appropriate setting for the min_kmer_cov parameter.
3.2. Variant calling and sequence variation in a de novo transcriptome assembly

In this study we set out to investigate the genetic variation present in 5 accessions of a *T. officinale* (common dandelion) apomictic lineage. To this end, a transcriptome assembly consisting of 122,456 transcripts produced by Trinity (highly similar to our original TrAll assembly) was used as a reference for our downstream analysis. This assembly was used for a differential gene expression analysis by J. Ferreira de Carvalho and we decided to use it for the remainder of our analysis in order to facilitate comparisons with their ongoing investigations.

Nowadays, variant calling can be carried out by wide range of different available software packages (for a review of the most commonly used ones see (Nielsen et al., 2011). For our analysis we chose the widely used GATK pipeline (Depristo et al., 2011). Apart from being considered as the golden standard for variant analysis, the GATK framework can handle RNA-seq data, at any level of ploidy. This goes through three major steps to finally report the called SNPs: i) the NGS reads are first mapped onto the reference assembly. The alignments produced at this stage are further processed through duplicate marking and per-base quality recalibration in order to assure the highest confidence in the next step, which is (ii) variant discovery. The tool used at this point, the HaplotypeCaller, can call SNPs in polyploid samples. (iii) Finally, (iii) the raw called variants are filtered, in order to distinguish between true polymorphic sites and machine artifacts.

3.2.1. Mapping to the reference assembly

The overall mapping percentage of the reads were similar for all 5 accessions (Table 6). More than 90% of the input reads were mapped (92.22%, 92.1%, 92.55%, 91.62% and 92.37% for accessions 3, 8, 11, 12, and 13, respectively), with the majority of these (more than 83%) being properly paired across all accessions. Less reads were available for accession 8, resulting in lower absolute values, but the mapping statistics are comparable with those of the other accessions. The percentages of singleton reads (reads for which of the mates was not mapped) were also similar and close to 2%.
### Table 6. Read mapping statistics on the reference de novo transcriptome assembly per accession.

<table>
<thead>
<tr>
<th>Input reads</th>
<th>3</th>
<th>8</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>223,205,072</td>
<td>190,258,354</td>
<td>230,041,130</td>
<td>234,320,418</td>
<td>242,178,302</td>
</tr>
<tr>
<td>Mapped (%)</td>
<td>205,845,661 (92.22)</td>
<td>175,222,600 (92.10)</td>
<td>212,901,342 (92.55)</td>
<td>214,678,066 (91.62)</td>
<td>223,706,146 (92.37)</td>
</tr>
<tr>
<td>Properly paired (%)</td>
<td>190,734,592 (85.45)</td>
<td>162,350,596 (85.33)</td>
<td>199,536,898 (86.74)</td>
<td>195,182,744 (83.30)</td>
<td>208,264,444 (86.00)</td>
</tr>
<tr>
<td>% Singletons</td>
<td>1.99</td>
<td>2.03</td>
<td>1.96</td>
<td>2.03</td>
<td>1.94</td>
</tr>
</tbody>
</table>

#### 3.2.2. Base score recalibration increases confidence in SNP calls

An important step of data preprocessing is the recalibration of the base scores of the mapped reads. The quality of the called SNP at a certain position depends both on the base calls in the reference and the per-base quality score of the reads mapping to that particular position (Nielsen et al., 2011). These quality scores are scaled versions of the probability of error in a certain SNP or base call. Since high-throughput sequencing methods are prone to generate errors in the short reads they produce, we need a way to ensure that the quality scores assigned to bases accurately reflect the true error probabilities. The GATK analysis pipeline comes with a tool that can recalibrate these scores (Base Recalibrator) in the context of different covariates, namely the sequencing machine cycle, the read group, the quality score and the trinucleotide context (Depristo et al., 2011). The algorithm in the Base Recalibrator uses the reported (raw) quality scores and for each covariate recalculates an empirical score, based on the observed mismatches in the input read alignments. The parameters of the mapping of reported to empirical quality scores can then be used to recalibrate any read set.

According to the online best practices of the GATK, this process leverages a database of known polymorphisms to distinguish read mapping errors from genotypic variation. In the absence of such a database, as in our case, after an initial round of de novo SNP calling a provisionalary database of such polymorphisms is produced, containing only the most confidently called SNPs. Subsequently, recalibration can be applied through multiple rounds of SNP calling, until convergence is reached. In contrast to well characterized genomes from model organisms (e.g. human), for which comprehensive SNP databases exist, the process of recalibration for non-model datasets is based on a trial-and-error approach. The exact number of iterations until the so-called convergence is reached cannot be known a priori. In addition, it can be rather computationally costly and much effort has to be put into its optimization.
The per base quality scores were recalibrated for each individual BAM file per accession in a total of 3 iterations and on a downsampled dataset, containing 0.01% of the total number of aligned reads. This resulted in the correction of the raw scores to more accurate empirical ones (Fig. 2).

![Recalibration curves showing the improvement in the per base quality scores before (pink) and after three rounds of recalibration (blue), for three different covariates: Reported quality score (A), machine cycle (B, D) and sequence context (C).](image)

**Figure 2.** Recalibration curves showing the improvement in the per base quality scores before (pink) and after three rounds of recalibration (blue), for three different covariates: Reported quality score (A), machine cycle (B, D) and sequence context (C).

### 3.2.3. Filtering called SNPs reduces their number to a high-confidence subset

Using the recalibrated data, a total of 387,740, 359,800, 382,847, 389,833 and 392,857 raw variants were called for accessions 3, 8, 11, 12 and 13, respectively (Table 7). This initial set of SNPs may contain several false positives, due to stochastic systemic sequencing and alignment errors.

**Table 7.** Number of raw SNPs called per accession before and after the application of different filtering steps.

<table>
<thead>
<tr>
<th></th>
<th>3</th>
<th>8</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Raw SNPs/Indels</strong></td>
<td>387,740</td>
<td>359,800</td>
<td>382,847</td>
<td>389,833</td>
<td>392,857</td>
</tr>
<tr>
<td><strong>SNPs</strong></td>
<td>350,016</td>
<td>325,463</td>
<td>345,216</td>
<td>351,944</td>
<td>354,216</td>
</tr>
<tr>
<td><strong>SNPs (Q100)</strong></td>
<td>278,840</td>
<td>253,955</td>
<td>272,926</td>
<td>281,207</td>
<td>283,839</td>
</tr>
<tr>
<td><strong>Final set</strong></td>
<td>165,285</td>
<td>149,385</td>
<td>159,079</td>
<td>167,045</td>
<td>168,080</td>
</tr>
</tbody>
</table>

In order to acquire a more reliable call set, we applied a series of filters. The raw calls contain predicted polymorphisms either at the level of one nucleotide or small insertions and deletions (indels). We first filtered out these small indels, based on the assumption that these will be more error prone. This resulted in the reduction of
the initial SNP set by approximately 10% for all accessions. The remaining SNPs were further filtered based on their estimated quality, as reported in the vcf files. We used a quality threshold of 100, that has been previously employed in similar analyses (Nicolaï et al., 2012). This further reduced the SNPs set by another 20%. Finally, following the online recommendations of the GATK pipeline for variant calling in RNA-seq data, we filtered out clusters of 3 SNPs within a window of 35 bases. This resulted in the final high quality SNP set that we used for the subsequent analysis and includes 165,285, 149,385, 159,079, 167,045 and 168,080 SNPs for accessions 3, 8, 11, 12, 13, respectively.

This yields a total of 218,761 SNPs (Fig. 3) of which, 120,684 are shared among all the accessions. Accession 13 showed the largest number of unique SNPs (10,379), while accession 8 had the lowest (5,910). Accession 8 showed lower levels of SNPs called throughout the whole analysis. This might be attributed to the lower number of sequencing reads, compared to the other accessions (Table 1, Table 7).

Figure 3. Venn diagram of the called SNPs for the 5 dandelion accessions.

We then wanted to see how the accessions are related with respect to their differences in SNPs. We first produced a table comprising all the SNPs and their presence or absence in all the accessions. Two difference measures were then used to calculate the distances between accessions, Euclidean and Jaccard. Finally we hierarchically clustered the accessions using the resulting matrices (Fig. 4). The results confirm that, based on the SNP calls, the 5 accessions are quite similar. The clustering patterns that we observe however, differ, depending on the distance measure used: with Euclidean distances accessions 8 and 11 cluster together, with
the accessions 3, 12 and 13 being more close (Fig. 4A), whereas using the Jaccard distance accession 8 seems to differentiate more than the other accessions that cluster together (Fig. 4B).

Based on the geographic origin of these accessions, we would expect that accessions 8 and 11 would cluster together, reflecting the same genetic variation, since they were both collected from the same field in Germany. Similarly, accessions 12 and 13 were both collected from a different field in Czech Republic. These expectations seem to be confirmed with the clustering using Euclidean distances, but not when using Jaccard distances.

Figure 4. Clustering of the dandelion accessions based on two different distance measures: Euclidean (A) and Jaccard (B).
4. Conclusions

This report details a first attempt to gain insight in the genetic variation within an apomictic *T. officinale* lineage, through a *de novo* transcriptome assembly. We were able to produce three different assemblies that showed what we believe is an overestimation of the number of predicted genes and transcripts. These discrepancies may be due to the inability of the assemblers to resolve differences caused by highly variable sequence information. This variability could be attributed to the presence of actual polymorphisms or errors. We attempted to improve these assemblies, by applying different tools, but our efforts were not successful. Further effort can be put into optimizing these tools, by applying different parameter settings specific to our needs. For example, using CD-HIT-EST with appropriate parameters could lead to better clustering of the transcripts based on sequence similarity, thus reducing their numbers. Such tools, though, come with a great range of parameters and their optimization can be time consuming.

Despite the aforementioned ambiguities in basic statistics, further post-assembly validation for the sequence content through homology-based comparisons, revealed that the transcriptome assemblies can be deemed complete, in terms of biologically functional sequences. Annotation of the content should be a priority for future work, since it cannot only help researchers assign function and identify potentially interesting genes, but it can help discriminate between computational artifacts and true, biologically meaningful sequences.

The variability in the reported results from each assembler, prompted us to examine in more detail how they perform on the transcriptome of two well characterized systems, the fission yeast and the budding yeast. This analysis was carried out in order to help us understand to a certain extent the way different parameters affect the resulting assemblies. It was carried out at a very basic level and its results confirmed that the *de novo* assemblers seem to overestimate the number of predicted genes. A more elaborate analysis taking into account more parameters, with a broader range of values would be needed in order to extract valuable conclusions. Furthermore, a post-assembly comparison of their various effects on the biological content, by means of e.g. BLAST comparisons with well annotated databases, would be also critical in understanding the inner workings of the assemblers. However, such an analysis was out of the scope of this study.

We also managed to quantify the variation between 5 members of an apomictic lineage in the form of SNPs. Our results seem to confirm the expectation that members of the same lineage share most of the genetic variation, and only a small fraction of the called SNPs is unique for each accession. The accessions were also clustered based on the calculated differences of present/absent SNPs, using two
different methods. Even though one of the methods is consistent with the biological reality, the clustering patterns differ, depending on the method.

This difference may be attributed to the confidence with which we distinguish present and absent SNPs in the accessions. In order to assure this confidence, efforts can focus on two directions. The one entails the application of a statistical test, like Fisher’s exact test, which can test the significance of SNP called absent (or present). The other one makes use of the coverage information that was originally used for calling the SNPs, to eliminate any SNP that was called at a position which is not covered at all in any of the other accessions.

Further efforts to understand the biological mechanisms underlying apomixis should foremost focus on assembling the genome of *T. officinale*. Research such as that presented here can benefit tremendously from the availability of a reference. Besides its use for validation of results on genetic variation, the availability of a reference genome can help to progress into research questions such as which alleles are present and expressed in different organisms, or even apomictic lineages. Second, a full annotation of the transcriptome in terms of biological functions and pathways and ranking of SNPs based on (predicted) functional effects would help researchers zoom in on the possible effects of the genotypic variation observed. Finally, to gain more insight into the role of epigenetic variation in these apomicts, it would be worthwhile to measure DNA methylation and histone modifications and correlate these to (lack of) genetic variation in specific genes, functions and/or pathways.
References


(Ranunculus auricomus complex, Ranunculaceae). Molecular Ecology 15, 897-910.


