Development of Genomic Resources for Ornamental Lilies (Lilium L.) *Arwa Shahin* *2012*

Development of Genomic Resources for Ornamental Lilies (Lilium L.)

Arwa Shahin

Development of Genomic Resources for Ornamental Lilies (*Lilium* L.)

Arwa Shahin

Thesis committee

Thesis supervisor

Prof. dr. R.G.F. Visser
Professor of Plant Breeding
Wageningen University

Thesis co-supervisors

Dr. ir. J. M. van Tuyl
Senior Scientist, Plant Research International
Wageningen University and Research Centre

Dr. P.F.P. Arens
Scientist, Plant Research International
Wageningen University and Research Centre

Other members

Prof. dr. ir. H.J. Bouwmeester, Wageningen University
Prof. dr. E.J. Woltering, Wageningen University
Dr. ir. S. Heimovaara, Royal van Zanten, Rijsenhout, the Netherlands
Dr. ir. S.A. Peters, Plant Research International, Wageningen

This research was conducted under the auspices of the Graduate School of Experimental Plant Science.

Development of Genomic Resources for Ornamental Lilies (*Lilium* L.)

Arwa Shahin

Thesis

submitted in fulfillment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof. dr. M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 19th June, 2012
at 4 p.m. in the Aula.

Shahin A.
Development of Genomic Resources for Ornamental Lilies (*Lilium* L.).
169 pages.

Thesis, Wageningen University, Wageningen, NL (2012)
With references, with summaries in Dutch and English

ISBN 978-94-6173-300-9

Table of contents

I dedicate this thesis to my beloved country ‘Syria’

Chapter 1

General Introduction

Ornamental plant industry

Ornamental production is economically an important sector. The worldwide production value of floricultural crops was estimated in 2006 to be around 50 billion euro, and the consumption value 100-150 billion Euro (Chandler and Tanaka, 2007). Flower consumption per person in Europe and Japan is comparable (28.7 and 26.5 US$ per capita respectively), and it is higher than in the United States (19 US$ per capita, Flower Council of Holland, 2007). Production of ornamental plants is divided into three main categories: cut flowers, pot plants, and garden plants. In 2010, the production of cut flowers in Europe was the highest with 56%, compared with 34% for pot plants and 9% for garden plants according to Dutch auctions (http://www.lei.dlo.nl/publicaties/PDF/2011/2011-029.pdf). In contrast, in the United States markets 70.5% of the production value was for garden and landscape flowers, 28% for pot plants, and only 1.5% for cut flowers (USDA, National Agricultural Statistics Service, www.nass.usda.gov).

The Netherlands is the heart of the international floriculture sector. Over 60% of world trade in flowers and plants takes place via Dutch auctions. The total output of horticulture in 2010 amounted to 7.9 billion Euros: 5.2 billion Euro comes from ornamental sector and 2.7 billion Euro from horticulture sector (December 2010, www.tuinbouw.nl). The most important cut flowers according to statistics of Dutch auctions are: tulip, *Chrysanthemum*, rose, gerbera, lily, and *Freesia* respectively (Fig. 1A); the most important potted plant are: *Phalaenopsis*, *Kalanchoe*, and hyacinth (Fig. 1B); and the most important garden plants are: *Viola, Pelargonium*, and *Osteospermum* (Fig. 1C), http://www.lei.dlo.nl/publicaties/PDF/2011/2011-029.pdf.

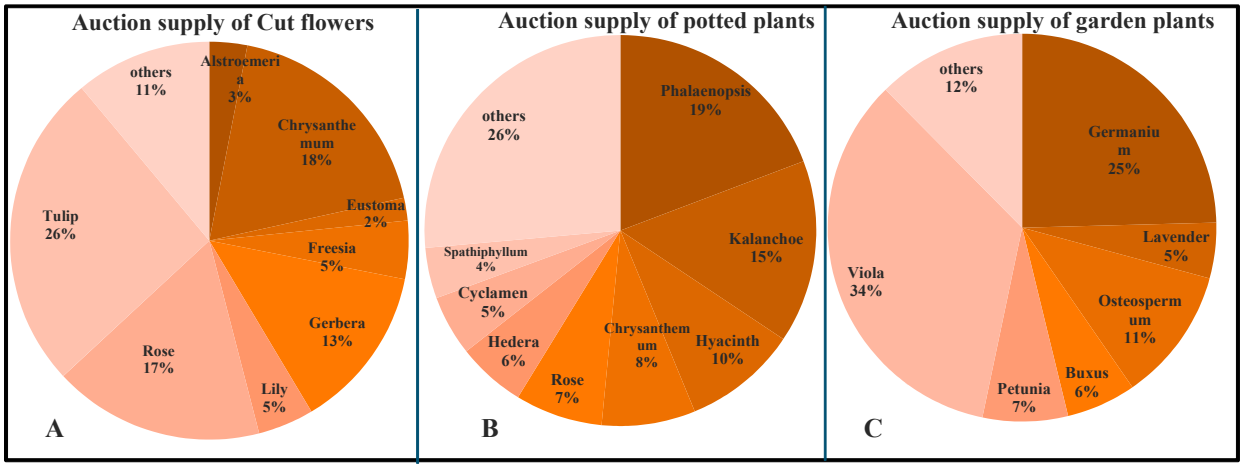


Figure 1. Statistics of the Netherland's Auctions supply of flowers in 2010: **A**, the supply of cut flowers, **B**, the supply of pot plants, and **C** the supply of garden plants.

Bulbous ornamentals

Ornamental bulb flowers or geophytes belong to more than 800 different genera. Production of geophytes is dominated by seven genera: *Tulipa* L., *Lilium* L., *Narcissus* L., *Gladiolus* L., *Hyacinthus* L., *Hippeastrum* Herb., and *Iris* L. that represent about 90% of the production area devoted to geophytes (Benschop et al., 2010). *Tulipa* and *Lilium* ranked as 3rd, 4th most important cut flowers at the flower auctions, while *Freesia* Klatt, *Alstroemeria* L., *Hippeastrum* Herb. and *Zantedeschia* Spreng. were, respectively, ranked as the 8th, 10th, 11th, and 13th most important cut flowers at the flower auctions (www.vbn.nl 2006). For many years, cultivation of flower bulbs was restricted to countries with developed ornamental horticulture industries and moderate climates such as the Netherlands. In the last two decades of the 20th century, production of bulb flowers in several other regions of the world has become important. For example, floricultural production in Latin America, Africa, and Asia is increasing rapidly. Cultivation of flower bulbs in the southern hemisphere is used for autumn flowering (October–December) in the northern hemisphere, especially in the United States, the Netherlands, Japan, Taiwan, China, and Canada (Benschop et al., 2010).

Figure 2. Some pictures of bulb flowers at the Keukenhof flower exhibition in the Netherlands (2011).

Flower bulbs are utilized for commercial bulb and flower production, including forced fresh-cut flowers, potted plants, and for landscaping, including private gardening (Benschop et al., 2010).

The best example of bulb utilizations is presented at the Keukenhof in the Netherlands. Over six million flower bulbs are on display as garden, pot, or cut flower (Fig. 2). Each spring, the breeders compete with each other in presenting, as part of their sales promotion, established favorites and the latest cultivars of spring-flowering bulbs (Bryan, 1989). The main production area of flowers in the Netherlands goes to bulbous flowers (Fig. 3).

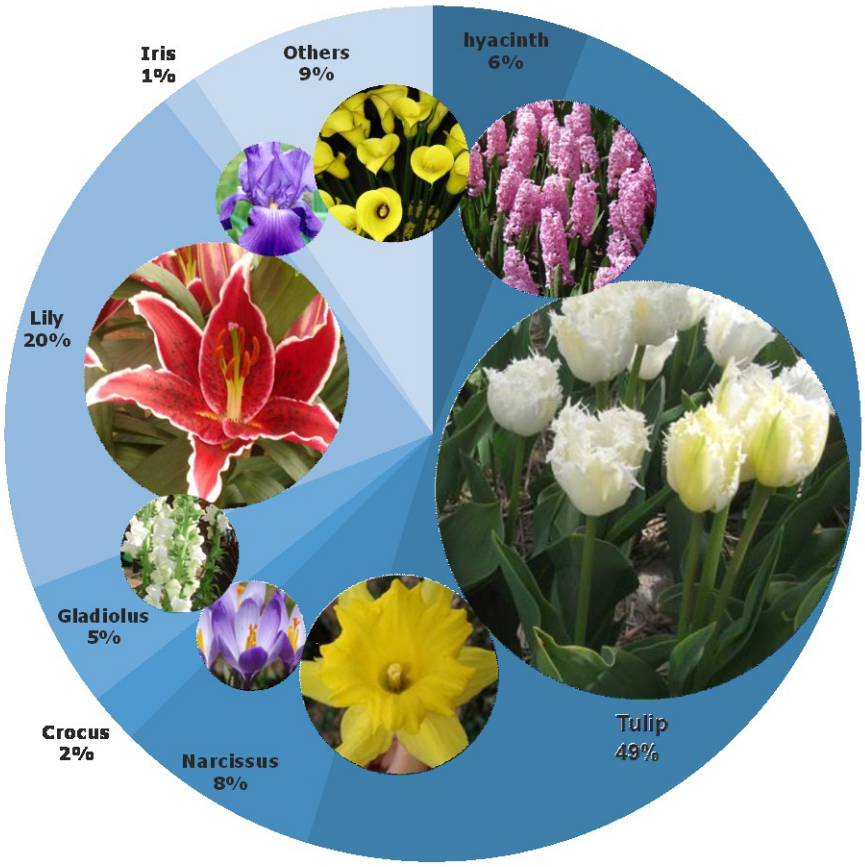


Figure 3. Estimated percentage of total propagation area of different flower bulbs by crop in the Netherlands, in 2010 (http://www.lei.dlo.nl/publicaties/PDF/2011/2011-029.pdf).

Breeding bulb flowers

Breeding objectives

The major objectives in breeding flower bulbs are: flower color, flower morphology, and plant architecture. Other increasingly important traits are: forcing time, yield, vase life, storability of the bulb, vulnerability of the flowers during transport, and disease resistance (*Fusarium*, *Botrytis*, and viruses), which are much more difficult to breed for since they are often polygenetic traits (Krens and Van Tuyl, 2011).

Conventional breeding

Conventionally, plant breeders recombine traits present in different parental lines of cultivated and/or wild species into single improved genotypes through various breeding schemes (Rahman et al., 2011). In flower bulbs, most of the breeding efforts have been carried out by private companies or hobbyists. In the past few decades, however, breeding has been performed more and more by specialized companies and research institutes (Benschop et al., 2010). The main strategies that were widely used for introducing new cultivars (variation) were: interspecific hybridization, spontaneous mutations, and polyploidization (Doorenbos, 1954).

Breeding strategy varies according to the genus, and it is affected by several factors: the availability of genetic diversity and the possibilities of using that variability, the length of the juvenile phase, and the propagation rate (Benschop et al., 2010; Krens and Van Tuyl, 2011). In *Gladiolus*, *Lilium*, and *Narcissus*, interspecific hybridization is the general way to produce new variation. However, interspecific hybridization in *Hyacinthus* is not possible since all genetic forms have been derived from one species with a range of different colors and ploidy levels (Doorenbos, 1954).

Successful interspecific hybridization depends to a great extent on the relationships within the genera; the closer the relationships the higher the success in crossings (Krens and Van Tuyl, 2011). For intersectional or intergeneric crosses pre- and post-fertilization barriers should be overcome. Pre-fertilization barriers are caused by pollen tube inhibition due to stigmatic incompatibility (Asano, 1980; Asano and Myodo, 1977a), which can be overcome by grafting or cutting style methods (Lim and Van Tuyl, 2006; Van Creij et al., 1997; Van Tuyl et al., 1991). Post-fertilization barriers resulting in seeds having no or just a little of endosperm and very small embryos which usually abort in early stage (Asano and Myodo, 1977b) can be overcome by applying *in vitro* techniques in the laboratory such as ovary culture or embryo rescue (Custers et al., 1995; Lim and Van Tuyl, 2006; Lu and Bridgen, 1996; Rhee et al., 2005; Van Tuyl and Van Creij, 2006).

Another strategy of introducing new variation in some traits is mutation breeding. Mutation breeding is based on the possibility of artificially inducing genetic changes in already existing cultivars (*e.g.* X-rays or gamma rays) or by exposing plant material to a number of chemical compounds known for their mutagenic properties. Also, spontaneous mutations for all kinds of traits do occur naturally (called sports). Often mutant cultivars can be grown and handled under more or less the same conditions as the original cultivar which is very convenient and economical for growers (Van Harten, 2002). In some crops, such as *Tulipa* and *Hyacinthus*, spontaneous mutations play an important role in cultivar diversification (Doorenbos, 1954). For example, among the 1,500 new tulip cultivars registered from 1996 to 2005, about 28% of them are mutants, and most have been spontaneous mutants.

Polyploidization has become very important in almost all flower bulb species. Currently, triploid and tetraploid cultivars predominate in *Narcissus* production. The role of polyploidy is important in hyacinths where the ploidy level varies from diploid, triploid to tetraploid. Also, in lily many cultivars >45% are polyploids (triploid, tetraploid, and few aneuploid). Polyploidization of cultivars is very interesting since it provides vigour to the cultivars: in general resulting in larger flowers and leaves (Van Tuyl et al., in press).

Polyploidization is also very important for further improving breeding process. Generally, after interspecific crosses, a series of back-crosses will be necessary to reduce the amount of donor DNA in the hybrid and to get back to the high quality of the original elite cultivars. In bulbous crops, F_1 hybrids obtained are often sterile (Krens and Van Tuyl, 2011). The fertility of F1 hybrids can be restored by doubling the chromosomes using colchicine or oryzalin 'somatic or mitotic polyploidization'. However due to the bivalent chromosome pairing during meiosis, somatic polyploidization does not produce intergenomic recombination which is essential for introgression of genes from one parent to the other (Lim and Van Tuyl, 2006; Ramanna and Jacobsen, 2003). Another way to achieve polyploidization is to use 2n-gametes that occur spontaneously in some genotypes, or that can be induced using for example laughing gas (N_2O) treatment (Barba-Gonzalez et al., 2006). This phenomenon occurs in lily hybrids, although at low frequencies and not in certain genotypes, and has been successfully applied in lily (Barba-Gonzalez et al., 2004; VanTuyl and Lim, 2003). Using this approach, recombination could be demonstrated and introgression led to new second generation hybrids combining new traits with existing quality (Barba-Gonzalez et al., 2005).

Various breeding strategies have been developed and applied and now there is a need for better selection to breed for the more difficult polygenic traits. In this context, molecular techniques (*in situ* DNA hybridization and molecular DNA markers) are potentially of great value for improving the efficiency of selection.

Marker assisted breeding

Conventional breeding in flower bulb is a slow process which can be speeded up by the use of molecular markers. In flower bulbs, molecular techniques have not been implemented yet, which might be due to the size of the genomes of the main genera. For example, the DNA content of the *Tulipa* genome (25-30 Gb) is about ten times the size of the human genome, and around 200 times higher than the DNA content of the model plant *Arabidopsis thaliana*.

Reasons for applying marker assisted breeding (MAB) can be that some traits are difficult to screen for or become scorable relatively late in a crop plant's life cycle such as disease resistance, abiotic stress tolerance, storability, flower color, or vase life. These traits become clear only when a plant is challenged by a pathogen, tested under specific conditions, or after a long juvenile phase that is needed for the plant to produce flowers (*e.g.* juvenile phase of tulip is five years and

several additional years are required for propagation). Additionally, some traits are polygenic or can be influenced by environmental conditions which make scoring of the traits is difficult.

Rapid and early screening of the progeny for specific traits can be achieved by developing molecular markers linked to these traits. This indirect selection by markers can significantly speed up the breeding process. Moreover, several traits can be easily observed concurrently. This methodology is widely applied in agricultural crops, such as maize or tomato (Agrama and Moussa, 1996; Martin et al., 1989). In bulbous ornamentals, molecular markers, genetic maps, and QTL (quantitative traits loci) analysis are still in their infancy. Because, molecular marker techniques also have important utility in plant breeding programs through assisting in plant variety protection as well as in distinctness, uniformity and stability testing (Heckenberger et al., 2006) their use is somewhat more established.

Molecular markers were developed for several bulb crops for the latter purposes. In Louisiana *Iris*, RAPD (random amplified polymorphic DNA) and retro-transposon markers (S-SAP) were developed for segregating populations resulting from interspecific hybrids (Kentner et al., 2003). Genetic maps and QTL analysis for several morphological, life history, and ecological traits were studied (Bouck et al., 2005; Bouck et al., 2007; Martin et al., 2007; Martin et al., 2008). Recently, around 400 EST-SSR (expressed sequence tags-simple sequence repeat) markers were developed for comparative genetic mapping and other genotyping applications in Louisiana *Iris* (Tang et al., 2009). In lily, primary genetic maps were constructed using AFLP (amplified fragments length polymorphism) markers in which *Fusarium* and virus resistance were identified (Van Heusden et al., 2002). In tulip, like in lily, a primary genetic map has been made using AFLP markers (Krens and Van Tuyl, 2011). In *Narcissus*, AFLP, RAPD, SSR, and NBS (nucleotide binding sites) markers were developed for genetic diversity, population genetic, and cultivar identification purposes (Lu et al., 2007; Simón et al., 2010; Wu et al., 2011). Similarly in *Crocus* L., retro-transposon markers, plastid sequences, RAPD, ISSR (inter-simple sequence repeats), and SSR markers were developed to analyze the genetic diversity and phylogenetic relationships in the genus (Alavi-Kia et al., 2008; Beiki et al., 2010; Rubio-Moraga et al., 2009). In *Alstroemeria*, genetic diversity was estimated using RAPD and AFLP markers (Aros et al., 2006; Han et al., 2000). In *Gladiolus*, molecular markers were developed for the first time in 2010 to estimate the genetic relationships among *Gladiolus* cultivars (Ranjan et al., 2010). In *Hyacinthus* and *Hippeastrum*, to our knowledge, no molecular marker work has been reported so far.

Molecular markers are relatively cheap, easy to produce and do not depend on environmental factors as compared to other types of markers. Nowadays, the rapid development in sequencing and genotyping technologies makes the generation and identification of molecular markers even faster, easier, and much cheaper (Krens and Van Tuyl, 2011).

Developing genomic and genetic resources for MAB

Molecular genetic resources for flower bulbs are very limited. Only few molecular markers have been developed, hardly any genetic maps are available, and the number of available EST in gene banks (if any) is very limited.

Next generation sequencing (NGS) technologies have dramatically reduced the time and the costs needed for sequencing. It now becomes feasible to generate large amounts of DNA sequence for species for which little or no prior sequence information exists, and to mine these sequences for polymorphisms (SNPs). The SNPs can be used to generate markers (SNP markers) for genotyping large numbers of individuals using high throughput SNP genotyping technologies. Once relevant number of markers becomes available for a certain species, these markers can be implemented to enhance the efficiency of conventional breeding. Molecular markers can be compared with the morphological observations of traits of interest to identify QTLs in both cross segregating populations and in association studies. In both cases, QTLs can be identified and implemented for MAB.

However, applying NGS in flower bulbs is not straightforward. The genomes of flower bulbs are large and rich with repetitive DNA which complicates the analysis (assembly) of sequence data. This constrain can be solved by genome complexity reduction methods. Another constrain is the need for expertise in bioinformatics in order to analyze and assemble the huge amount of sequence data generated using NGS technologies. In out crossing ornamental bulbs, assembly become more challenging due to: the lack of genome sequence support of the gene banks that can be used as a backbone for assembly step, and the highly polymorphic nature of the out crossing flower bulb genome. Many assembler programs are currently used to assemble the NGS data, but little is known about assembler's performance, especially when dealing with highly heterogeneous species and how that influences SNP retrieval. This field has become a new area of research.

Developing genomic resources for plant systematics

Traditionally, chloroplast DNA (cpDNA) and nuclear DNA (ribosomal gene spacers) were widely used for plant systematic studies (Small et al., 2004). Very few different gene sequences were used in this kind of studies. For instance, cpDNA and ITS were used extensively to classify *Lilium* species and cultivars (Dubouzet and Shinoda, 1999; Muratović et al., 2010; Nishikawa et al., 2001; Nishikawa et al., 1999; Rešetnik et al., 2007). With the vast developments in NGS technologies, huge amounts of sequence data are becoming available that can be used for plant systematic studies. Nuclear DNA sequences have the advantage that: their evolutionary rate is higher than in plastid DNA, the inheritance is bi-parental, and there is an abundance of long and independent genes (Small et al. 2004). Furthermore, the ability to identify heterozygosity within hybrids (allelic variation) can give better estimations of phylogenetic relations because twice the amount of information is provided (Joly and Bruneau, 2006; Liu et al., 2008). However, the

application of NGS in plant systematic studies is still limited since a full spectrum of tools to build species phylogenies form, is not well established yet (Sanderson and McMahon, 2007).

Nevertheless, some interesting hypothesis, for instance whether breeding/domestication processes are being imprinted in a species genome, can be best tested using NGS data. This can be measured by looking for positive selection at the nucleotide/codon level. If the ratio of functional to non-functional change (omega value) at the codon-level is > 1 that indicates the possibility of positive selection, while if the ratio is < 1 that will indicate a purifying selection (Yang and Dos Reis, 2011). This topic is very recent and is not yet applied on ornamental plants.

Ornamental traits desirable to improve

There are several important traits that the producers and consumers wish to have improved in ornamental plants such as: diseases resistances (*Fusarium*, *Botrytis*, and viruses), postharvest quality, flower color, and scent. The MAB is commonly used approach to understand genetic background and to speed up the breeding process in several crops. So far, few studies focused on identifying the genetic background of such traits, like *Fusarium oxysporum* and Lily Mottle Virus (LMoV) resistance in *Lilium* (Shahin et al., 2009; Straathof and Loffler, 1994; Van Heusden et al., 2002), TBV (tulip breaking virus) resistance in *Tulipa* (Krens and Van Tuyl, 2011), flower color and flower spots in *Lilium* (Abe et al., 2002; Nakano et al., 2005; Van der Meulen et al., 1996). Alternatively, genetic modification approach which is less accepted by public has been applied to some extend in flower bulbs. Some studies were carried out, in bulbous flowers, to establish genetic transformation protocols such as in *Alstroemeria, Gladiolus*, *Lilium*, and *Tulipa* (Akutsu et al., 2004; Chauvin et al., 1997; Kamo et al., 1995; Kamo and Han, 2008). In *Hyacinthus, Fusarium* resistance was successfully transferred to hyacinth varieties (Popowich et al., 2007). However, this is only possible when: a good protocol for transformation is well established, the trait of interest is well studied, and the genes responsible for this trait are precisely identified.

For some traits, genetic improvement is still challenging such as for flower longevity. Flower longevity is an important characteristic for ornamental plants. In some ornamental like in carnation, which is ethylene-sensitive, vase life was improved by producing transgenic carnation in which ethylene biosynthesis or ethylene perception was altered (Chandler, 2007). However, this is more complicated for species that are ethylene-insensitive like *Lilium* and *Tulipa* since the factors that regulate their vase life are not defined yet. QTL mapping is not straightforward approach for vase life since a whole population (progenies) has to be tested for their vase life with replicates (Van der Meulen et al., 1996) which is very laborious and expensive. Vase life is a very complex trait as both flowers and leaves senescence are involved and different responses can occur. Therefore, splitting up this trait in different compounds and first identifying the main

drivers for single compound will enable to simplify the complex trait into sub-trait that are amenable for studying and useing in selection for breeding.

Lilium L.

Species of genus *Lilium* originate from Asia, Europe, and North America (Bryan, 1989) are mostly vegetative propagated monocot perennials and are one of the economically most important flower bulbs. The genus *Lilium* (*Liliaceae* family) comprises around 100 species and more than 9,400 cultivars (International Lily register, http://www.lilyregister.com/). The species of this genus were taxonomically classified into seven sections based on 13 morphological and two germination characteristics. The seven sections are *Martagon, Pseudolirium, Lilium (Liriotypus), Archelirion, Sinomartagon, Leucolirion*, and *Oxypetalum* (Comber, 1949; De Jong, 1974).

In general, wild species within each section are relatively easy to cross and the hybrids are fertile (McRae, 1990; Van Tuyl et al., 2002). The interspecific hybrids within the sections especially those within the sections *Leucolirion*, *Archelirion*, and *Sinomartagon* represent the most important breeding groups which are:

1. *Longiflorum* hybrids (L genome). They originate from intra- or inter-specific hybridization with *L. formosanum* in the *Leucolirion* section, have trumpet-shaped, pure white flowers, a distinctive fragrance, year-round forcing ability and mostly outward-facing flowers (McRae, 1990).
2. Asiatic hybrids (A genome). They are derived from interspecific crosses among at least 12 species of the *Sinomartagon* section (Leslie, 1982). Their cultivation can be traced to the early 1800s in Japan (Shimizu, 1987). Cultivars of Asiatic hybrid lily have a wide color-variation in their tepals (orange, white, yellow, pink, red, purple, and salmon) and early to late flowering (Woodcock and Stearn, 1950). Some species in this section show resistance to *Fusarium* and viruses (McRae, 1998a).
3. Oriental hybrids (O genome). They result from hybridization among five species of the *Archelirion* section. Generally, Oriental hybrids are late-flowering, with big and showy flowers with a pleasant fragrance (McRae, 1998a). Most species are resistant to *Botrytis elliptica* that affects most of the lilies from the other sections (Barba-Gonzalez et al., 2005).

Lilies have a wide variety of valuable characters such as flower size, color, flowering time, and resistance to different pathogens. Combining these vital horticultural traits into one cultivar by crossing is almost the only way to obtain introgression of traits, since genetic transformation approaches are not well developed for lily yet. Possibilities for cross combinations in *Lilium* between the species of the seven sections are limited by incompatibility and incongruity which are due to: pre-fertilization and post-fertilization barriers. To overcome these barriers, integrated

methods such as grafted-style, *in vitro* pollination, embryo rescue, and ovule culture techniques are needed (Van Tuyl et al., 1991). Using these methods, many lily interspecific hybrids have successfully been made. For instance, *L. longiflorum* (*Leucolirion*) x *L. rubellum* (*Pseudolirium* section), *L. longiflorum* x *L. candidum* (*Lilium* section), *L. longiflorum* x Asiatic hybrids (*Sinomartagon*) (Van Tuyl et al., 2000). However, most of these inter-specific hybrids tend to be sterile (Van Tuyl et al., 2002). Chromosome doubling and *2n* gametes (gametes with somatic chromosome numbers) (Ramanna and Jacobsen, 2003) have been used to restore the fertility of inter-specific hybrids in lily.

Lilium species have been extensively used for cytological investigation. Basic studies on chromosome identification and karyotype analysis (Stewart, 1947) were conducted. The newer molecular cytogenetic techniques such as FISH (Fluorescent *in situ* hybridization) and GISH (Genomic *in situ* hybridization) have enabled researchers to investigate the meiosis and the homoeology of *Lilium* in detail. The restitution mechanisms that lead to unusual chromosome constitution in *2n* gametes have been revealed by GISH (Karlov, 1999; Lim, 2000). Barba-Gonzalez (2005) studied the occurrence of *2n* gametes in the F1 hybrids of Oriental × Asiatic lilies and used them for production of sexual polyploids from sterile Oriental × Asiatic hybrids. One of the most important advances has been achieved with the development of diploid backcross progenies (BC1 and BC2) from interspecific *Longiflorum* x Asiatic hybrids backcrossed to Asiatic cultivars (Khan, 2009; Zhou, 2007). Additionally, meiosis of interspecific hybrids was followed and cytological maps of three complete genomes of lilies (L, A, O) based on the recombination sites in the BC progeny of two interspecific hybrids (Khan et al., 2009) were constructed.

On the other hand, genetic mapping of lily has not yet been well studied. So far, RAPD and ISSR markers were used to construct genetic linkage maps and to map anthocyanin and carotenoid pigmentations in the progeny of ‘Montreux’ x ‘Connecticut King’ (Asiatic hybrids) (Abe et al., 2002; Nakano et al., 2005). Additionally, RAPD and AFLP markers were used to construct genetic linkage maps and to map *Fusarium* resistance using progeny of ‘Connecticut King’ x ‘Orlito’ (Asiatic hybrids) (Straathof et al., 1996; Van Heusden et al., 2002). Those genetic maps are far from saturation and the marker types difficult to be converted into simple PCR markers. Therefore, more genetic markers need to be added to the map, preferable of a type that is universal.

This thesis deals with the establishment of novel genomic resources for molecular and other genetic studies in lilies.

Scope of this thesis

We studied the genetics and physiology of lily with the aim to improve the efficiency of breeding and selection in this crop. For that, we build genetic resources using NGS technology. A large set of EST sequences were generated and annotated, SSR markers were mined, SNP markers were developed and used for genotyping two mapping populations of lily, and genetic linkage maps were constructed. Also, these resources were implemented to conduct comparative genomic studies between *Lilium* and *Tulipa* to estimate the genetic distances, and to detect if there is a signature of positive selection occurred in lily and tulip genomes. Additionally, physiology of vase life of lily flowers was studies to identify possible regulator(s) of lily flower longevity.

Chapter 2 presents the construction of genetic linkage maps for two populations, LA (*L. longiflorum* 'White Fox' x Asiatic 'Connecticut King') and AA ('Connecticut King' x 'Orlito'), using AFLP, DArT, and NBS profiling markers. Common markers were identified and used in aligning the linkage groups of the two populations. Several horticultural traits (flower color, stem color, flower spots, flower direction, and antherless phenotype) were mapped. Six putative QTLs for *Fusarium oxysporum* resistance were identified and mapped, and LMoV resistance was mapped as a locus on AA maps.

In chapters 3, 4 and 5 we presented the generation and implementation, for the first time, of ESTs data produced by 454 pyro-sequening for MAB applications in *Lilium* and *Tulipa*. In **Chapter 3**, NGS data from a highly polymorphic outcrossing species 'lily' were analyzed. All the steps that are needed for SNP markers retrieval were discussed. Two different assembler programs (CAP3 and CLC) that use two different approaches (OLC: overlap layout consensus and De Bruijn algorithms) for data assembly were compared. In **Chapter 4**, cDNA sequence data generated of *Lilium* and *Tulipa* were descriped. Transcriptomes of four lily genotypes ('Connecticut King', 'White Fox', 'Star Gazer', and 'Trumpet') and of five tulip genotypes ('Cantata', 'Princeps', 'Kees Nelis', 'Ile de France', and 'Bellona') were sequenced using 454 pyro-sequencing and assembled using CLC assembler. The SNP and SSR markers were mined for both genera. Transcriptomes of *Lilium* and *Tulipa* in addition to the orthologous genes identified between both of them were annotated and described according to Gene Ontology terminology. In **Chapter 5**, SNP markers were used for genotyping two half-sib mapping populations of lily (LA and AA populations) using KASP technology. The successfully generated SNP markers together with the AFLP, NBS, and DArT markers were used to re-construct the genetic maps of these two populations.

In **Chapter 6,** a comparative genomic analysis among the four sequenced genotypes of *Lilium*, the five sequenced genotypes of *Tulipa*, and between the two genera using a total of 47 gene contigs was conducted. These data were used to establish a straightforward and simple methodology to use allelic nuclear sequence data generated using RNA-seq technology to

estimate the genetic divergence among genotypes. The positive selection (functional changes at codon-level, or omega value >1) was calculated for each gene contig and compared with genetic distances of those genes.

In **Chapter 7**, vase life of seven lily genotypes was investigated to identify the factor(s) that control vase life in lily. We compared the vase life of lily cultivars under two treatments: tap water with HQS (8-HydroxyQuinolinol Sulfate), and tap water with sugar and HQS. The effect of sugar treatment on: vase life, dry weight, and abscisic acid concentration in flowers were studied. The concentration of ABA at anthesis and senescence were measured using LC-MS-MS.

Finally in **general discussion** (Chapter 8) the results of preceding chapters were evaluated in a wider perspective and their implications for improving breeding programs were discussed.

Chapter 2

Genetic Mapping in *Lilium* L.: Mapping of Major Genes and QTLs for Several Ornamental Traits and Disease Resistances

Arwa Shahin [1,2], Paul Arens [1], Adriaan W. van Heusden[1], Gerard van der Linden[1], Martijn van Kaauwen[1], Nadeem Khan [1], Henk J. Schouten[1], Eric van de Weg[1], Richard G. F. Visser[1] and Jaap M. Van Tuyl[1]

[1] Wageningen UR Plant Breeding, Wageningen University and Research Centre, P.O. Box 386, 6700 AJ Wageningen, the Netherlands

[2]Graduate School Experimental Plant Sciences, Wageningen University

This chapter has been published with modification in Plant Breeding (2011) 130:372-382

Abstract

Construction of genetic linkage maps for lily was achieved using two populations, LA and AA that share one parent 'Connecticut King'. Three different molecular marker systems (AFLP™, DArT, and NBS profiling) were used in generating linkage maps for 'Connecticut King'. The LA and the AA populations consist of 20 and 21 linkage groups, respectively. Average density between markers was 3.9 cM for the LA, and 5 cM for the AA population. Several horticultural traits were mapped for the first time in *Lilium* and showed to be single gene based. We propose to name these genes as *LFCc* for flower color, *lfs* for flower spots, *LSC* for stem color, *lal* for antherless phenotype, and *lfd* for flower direction whereby upper and lower case names refer to dominant and recessive genes, respectively. Additionally, LMoV was mapped as a locus on linkage group AA10. For *Fusarium* resistance, the Kruskal-Wallis test identified six putative QTLs in AA population of which one QTL (explaining 25% of the variation in resistance) could be confirmed by interval mapping.

Introduction

Genus *Lilium* belongs to *Liliaceae* family and comprises more than 80 species (Asano, 1989). During the last 50 years, more than seven thousand cultivars have been developed (International Lily register, http://www.lilyregister.com/). These are classified mainly into three groups: Longiflorum (L), Asiatic (A), and Oriental (O) hybrids that belong to different taxonomic sections: *Leucolirion*, *Sinomartagon*, and *Archelirion* respectively. Longiflorum hybrids have trumpet shaped white flowers that are mostly side-facing and have a distinctive fragrance. They show strong growth vigour and year round forcing abilities (McRae, 1998b). Asiatic hybrids are important due to their wide variation in flower color and shape, early flowering, and resistance to *Fusarium* and lily mottle virus (LMoV). Oriental hybrids have large and attractive flowers with a wide range of white, pink, and yellow colors, strong fragrance and resistance to *Botrytis elliptica*. Crosses within a group are relatively easy but crosses between species or cultivars of different groups are very difficult due to pre-fertilization (pollen tube inhibition) (Asano and Myodo, 1977a) and post-fertilization barriers (endosperm degeneration and embryo abortion) (Asano and Myodo, 1977). Applying cut style and embryo rescue techniques allow overcoming these pre- and post-fertilization barriers (Van Tuyl et al., 1991).

In breeding, molecular linkage maps are important for quantitative and qualitative trait analysis. However, despite the economic importance of lily, only two genetic mapping studies have been published up to now. This may be due to the fact that lily species (2n=2x=24) have one of the largest genomes (approximately 36 Gbp, http://www.rbgkew.org.uk/) among the plant and animal kingdoms. For instance, the genome size is estimated to be more than 200 times larger than that of *Arabidopsis thaliana*. The difficulties in markers analysis due to the huge genome size

together with a long generation time of three years and the heterozygous genome structure of lilies might have formed a considerable barrier for the execution of mapping studies.

The currently published linkage maps can be regarded as preliminary because relatively few markers have been used and many small linkage groups remained. Abe et al. (2002) used RAPD (randomly amplified polymorphic DNA) and ISSR (inter-simple sequence repeat) markers to construct parental linkage maps with 95 and 119 markers respectively in a cross between the two Asiatic cultivars 'Montreux' and 'Connecticut King' to elucidate the genetics of floral anthocyanin pigmentation. Van Heusden et al. (2002) used AFLP^{TM} markers in an Asiatic backcross population to map disease resistance against two important diseases: *Fusarium oxysporum* and LMoV. Four QTLs for *Fusarium* resistance were identified whereas LMoV resistance was controlled by a single locus. Only the linkage group bearing the LMoV locus was published (Van Heusden et al., 2002). Even though a common parent, cultivar 'Connecticut King', was used to construct the genetic maps in these two studies, a comparison of the genetic linkage maps was not possible due to the different types of markers used.

For accurate trait mapping and initiating marker assisted breeding, good marker coverage of maps is required, preferably with markers that target the traits of interest and/or can be easily used in downstream breeding applications. Therefore, two relatively new molecular marker techniques were applied on lily. The DArT (Diversity Arrays Technology) is a high throughput molecular marker system (Mace et al., 2008) from which markers are expected to be more readily converted. The NBS profiling was used because it targets resistant genes and resistance gene analogs (Van der Linden et al., 2004). The aims of this study were: 1) To test the feasibility of DArT and NBS profiling in lily, 2) To develop higher density genetic linkage maps for two lily populations, one of which is a continuation of the mapping study of Van Heusden et al. (2002), 3) To map important ornamental traits which segregate in these two lily populations and re-map LMoV and *Fusarium* resistance on the AA maps.

Material and Methods

Plant material

Two mapping populations were used. The first is the newly generated LA population, which is a F1 population of 98 genotypes made in 2000 from a cross between Longiflorum 'White Fox' x Asiatic 'Connecticut King' using cut style pollination and embryo rescue. The second is an AA population of 100 individuals (Straathof et al., 1996; Van Heusden et al., 2002), which was produced in 1989. It is a backcross of 'Connecticut King' with 'Orlito' (from the cross 'Connecticut King' x 'Pirate'). Cultivar 'Connecticut King', which is the common parent in both populations, is a well-known Asiatic cultivar resistant to LMoV and partially resistant to *Fusarium oxysporum*. It has yellow, spotless flowers and green stem color. The Longiflorum parent 'White Fox' is susceptible for both diseases, has white, spotless, out-facing flowers and

green stem color. 'Pirate' is susceptible to *Fusarium* and LMoV, has orange flowers with spots and a dark green stem color. 'Orlito' is partially resistant to *Fusarium* and susceptible to LMoV, has orange flowers with few spots and dark green stem color (Table 1).

Table 1: the segregation of the mapped traits of the AA and LA populations

	Connecticut King	White Fox	Orlito	LA population		AA population	
Flower color	yellow	white	orange	all are white	--	42yellow:55 orange	χ^2= 1.74
Spots	spotless	spotless	with spots	61 spotless:29 with spots	χ^2=2.5	51spotless:46 with spots	χ^2=0.09
Stem color	green	green	dark green	77 green:15 dark green	χ^2= 3.7	53green:46 dark green	χ^2= 0.5
Antherless	with anther	with anther	with anther	all have anthers	--	77with anthers:20 antherless	χ^2=1.01
Flower direction	up-face	out face	up face	63 out-face:27upface	χ^2=1.2	all are up face	--
Fusarium	resistant	susceptible	partial resistant	disease test still running	--		
LMoV	resistant	susceptible	susceptible	disease test still running	--		

DNA extraction

Young leaves were collected from all plants. Leaves were frozen in liquid nitrogen and stored at -80° C upon DNA isolation. DNA was isolated using the either the Qiagen DNAeasy extraction kit for NBS profiling, or the protocol of Fulton et al. (1995) for the DArT analysis.

Molecular markers

New NBS and DArT marker data were developed for both populations. Additionally, for AA already existing AFLP data (Van Heusden et al. 2002) were included.

NBS Profiling

NBS profiling has been performed according to Van der Linden et al. (2004). In brief, 400 ng DNA of each plant was digested for 4 h with one of the restriction enzymes *Mse*I, *Alu*I, *Taq*I, *Hae*III, *Rsa*I, and *Msp*I (*Msp*I is a methylation sensitive restriction enzyme only used with NBS3) followed by heat inactivation. A special adapter was ligated to the fragments and NBS-specific fragments were amplified involving a two-step PCR amplification with one of the NBS primers and an adapter specific primer (Table 2). The radioactively labelled PCR products were separated on a 6% polyacrylamide gel, and visualized by autoradiography. Polymorphic bands were scored as present or absent. In LA population, due to experimental problems (initial DNA quality), only three restriction enzymes (*Mse*I, *Hae*III, and *Rsa*I) with primer NBS6 were scored.

NBS markers were named according to the names of the two NBS primers used (NBS3 or NBS6), followed by the first letter of the restriction enzyme used and a number reflecting the fragment order on the gel.

Table 2: adapter and primer sequences used for generation of NBS profiling and DArT markers in *Lilium*

	Sequence 5'-3'
Adapter sequence (DArT)	GTT CAG TCA TAG ATG GTG CA CCA TCT AAC TTG ACT G
Adapter primer (DArT)	CAG TCA AGT TAG ATG GTG CAG
NBS3	GTWGTYTTICCYRAICCISSCATICC
NBS6	YYTKRTHGTMITKGATGATATITGG
NBS blunt adaptor (long arm)	ACTCGATTCTCAACCCGAAAGTATAGATCCCA
NBS blunt adaptor (short arm)	TGGGATCTATACTT
Primer adapter (NBS)	GTTTACTCGATTCTCAACCCGAAAG

DArT markers

- Array construction

Diversity array technology (DArT) was developed in lily according to (Jaccoud et al., 2001). Because resistance to *Fusarium* and LMoV is one of the most important breeding goals for MAB in lily, and because cultivar 'Connecticut King' is the sole source of resistance in these populations, DNA from this cultivar was used for construction of DArT micro-arrays. Briefly, genomic DNA of 'Connecticut King' (250 ng) was digested with the methylation sensitive restriction endonuclease *Pst*I, and an adapter (Table 2) was ligated to these fragments. An extra digestion with a frequent cutter enzyme was applied in order to reduce the number of fragments that potentially can be amplified (optimal number of fragments is 10.000 to 15.000; Andrzej Kilian, Diversity Arrays Technology Pty Ltd, pers. comm.). Only non-digested *Pst*I-*Pst*I fragments with adapters at both sides can be amplified. *Pst*I- amplicons of 'Connecticut King' were ligated into the PCR2.1-TOPO vector using the TOPO cloning kit and electroporated into *Escherichia coli* cells (TOP10F Invitrogen). The M13-universal primers were used for amplification and the products were spotted in triplicate on single poly-L-lysine-coated slides (Erie Scientific Com) using a MicroGrid II micoarrayer (Biorobotics, UK). The size of the spotted fragments ranged between 500- 2000 bp (Khan, 2009). A set of polymorphic clones showing segregation in AA population was also used for detecting polymorphism in LA population.

- Genotyping

DNA from the individuals of LA population was digested by *Pst*I/*Taq*I whereas also *Pst*I/*Mse*I was used in AA population. *Pst*I adapters were ligated to the restricted fragments. The *Pst*I/*Pst*I fragments were amplified using the adapter primer (Table 2). For each individual approximately 650 ng PCR product was labelled with fluorescent Cy5-dUTP or Cy3-dUTP (Pharmacia) using random decamers and DNA Polymerase I (NEB/Fermantas) and then hybridized to the arrays (one genotype per array). Arrays were scanned using a LASER micro array scanner (ScanArray Express HT Microarray Scanner) and images were generated for each of the fluorescent dyes using the appropriate laser/filter combination for Cy-3 and Cy-5. DArTsoft, a software package developed by DArT P/L (http://www.diversityarrays.com/ index.html) was used to automatically

analyze each batch of TIF image pairs generated. Only markers where the probability of belonging to one of the two possible clusters (present/absent), based on normalised hybridisation signals, was above 0.95 were incorporated into a 0/1 scoring table (see http://www.diversityarrays.com/molecularprincip.html).

DArT markers in LA population were named according to the plate number and well position of the *E. coli* colony in the corresponding microtiter plate, followed by a number that refers to the order of the markers according to the *P* value for scorability (Wittenberg et al. 2005). DArT markers in AA population were named either as LPM (lily *Pst*/*Mse*), or as LPT (lily *Pst*/*Taq*).

Phenotypic data

Resistance to *Fusarium* and/or LMoV was scored as previously described (Van Heusden et al., 2002). In brief, *Fusarium* resistance was tested for four years (1992, 1993, 1994, and 1999). Six to ten weeks after planting in artificially infected soil, bulblets were visually inspected and a numerical classification was given: 1= healthy; 2=slightly rotten; 3=moderately rotten; 4=heavily rotten; 5=very heavily rotten; and 6=completely decayed.

Additionally, several horticultural traits that segregat in these two populations were scored (Table 1). In AA population, flower color varied between the basal and upper part of the flower and therefore this trait was scored separately for these two parts as orange or yellow. Flower color in LA population did not segregate (Table 1). Spots (the number of spots on the petals) and stem color (green, dark green) segregated and were scored in both populations. Antherless segregated in AA population and was scored as presence/absence of anthers. In LA population flower direction segregated and was scored as outfacing/upfacing (Table 1). Segregations ratio of phenotypic traits were tested using the Chi-square test with a significance threshold of P=0.05.

Map construction

Genetic maps were constructed using JoinMap® 4.0 (Van Ooijen, 2006) with Haldane's mapping function, linkage with a recombination threshold of 0.45, and a LOD threshold of 1. Markers with identical scores were reduced to one marker only prior to grouping. Grouping was based on independence LOD (with LOD > 4) using the regression method. Sets of markers, that mapped around the same position and differed for just one or two missing values, were excluded but for one representative. The segregation ratio of alleles for each locus was evaluated by Chi-square test with a significance threshold of P=0.05. The expected segregation ratios were 1:1 or 3:1. In LA population only the <nnxnp> marker type was used in mapping, while both <lmxll> and <hkxhk> marker types were used for mapping in the backcross AA population.

Marker scores were checked for putative double recombination events using the genotype probabilities output from JoinMap®. The indicated markers were checked back only for NBS and AFLP markers (few scoring or typing errors were corrected), since DArT markers were scored using computer software and a high probability threshold for inclusion as marker. Similarly,

phenotypic traits were checked back for all the genotypes that showed double recombination events. Linkage groups with a mean Chi-square above 2 were further analyzed to identify markers causing tension in the map. In a number of cases, markers that caused high tension were omitted from the final map. MapChart (Voorrips, 2002) was used to draw the genetic linkage maps.

QTL mapping

Since genome coverage and marker density in AA population was increased compared to by Van Heusden et al. (2002), disease resistance against LMoV and *Fusarium* was re-mapped using MapQTL 5.0 (Van Ooijen et al., 2004) and the resistance scores data of Van Heusden et al. (2002). In short, the nonparametric Kruskal-Wallis test was used to initially map *Fusarium* resistance OTLs using the results of separate disease tests from four years (1992, 1993, 1994 and 1999). Furthermore, interval mapping and rMQM were also applied to confirm the putative QTLs identified by Kruskal-Wallis. For this ^{10}Log transformation was performed to obtain normally distributed data. A permutation test with 1000 replications (Churchill and Doerge, 1994) was carried out to establish the LOD threshold. QTLs detected by interval mapping were used as cofactors for rMQM until the number of QTL became stable. This procedure was also applied for mapping the number of spots.

Results

NBS markers

NBS profiling generated 15-60 scorable polymorphic markers per NBS-primer. Amplification with NBS6 primer consistently produced more markers than with NBS3, the best primer/enzyme combination was NBS6/*Mse*I which resulted in 60 markers in AA population. In LA population, 34 <nnxnp> markers segregating from 'Connecticut King' were scored. In AA population, in total 53 <lmxll> and 102 <hkxhk> markers were scored. Since 'Connecticut King' was used in both populations, a number of NBS markers were detected as common markers (13 NBS makers). Eight of them were mapped in both populations and were named Com.NBS-1 to 8 (Fig. 1).

DArT markers

In LA population, 687 polymorphic DArT markers were identified. Thirty four of them were not used in mapping because they were either specific for 'White Fox' or they were of <hkxhk> type. In AA population, a lower number of DArT markers (338 in total) were polymorphic as can be expected from a backcross. A total of 62 <lmxll> and 182 <hkxhk> DArT markers were used in mapping, while the remaining 94 were of <nnxnp> type. The polymorphic DArT clones from AA population (<lmxll>) were also used for the LA population to generate common DArT markers. Unfortunately, the number of markers that was polymorphic in LA population as well was limited

(21 markers), however they were useful in combining and connecting some of the linkage groups of the two crosses. The common DArT markers were coded as (Com.DArT- 1 to 21).

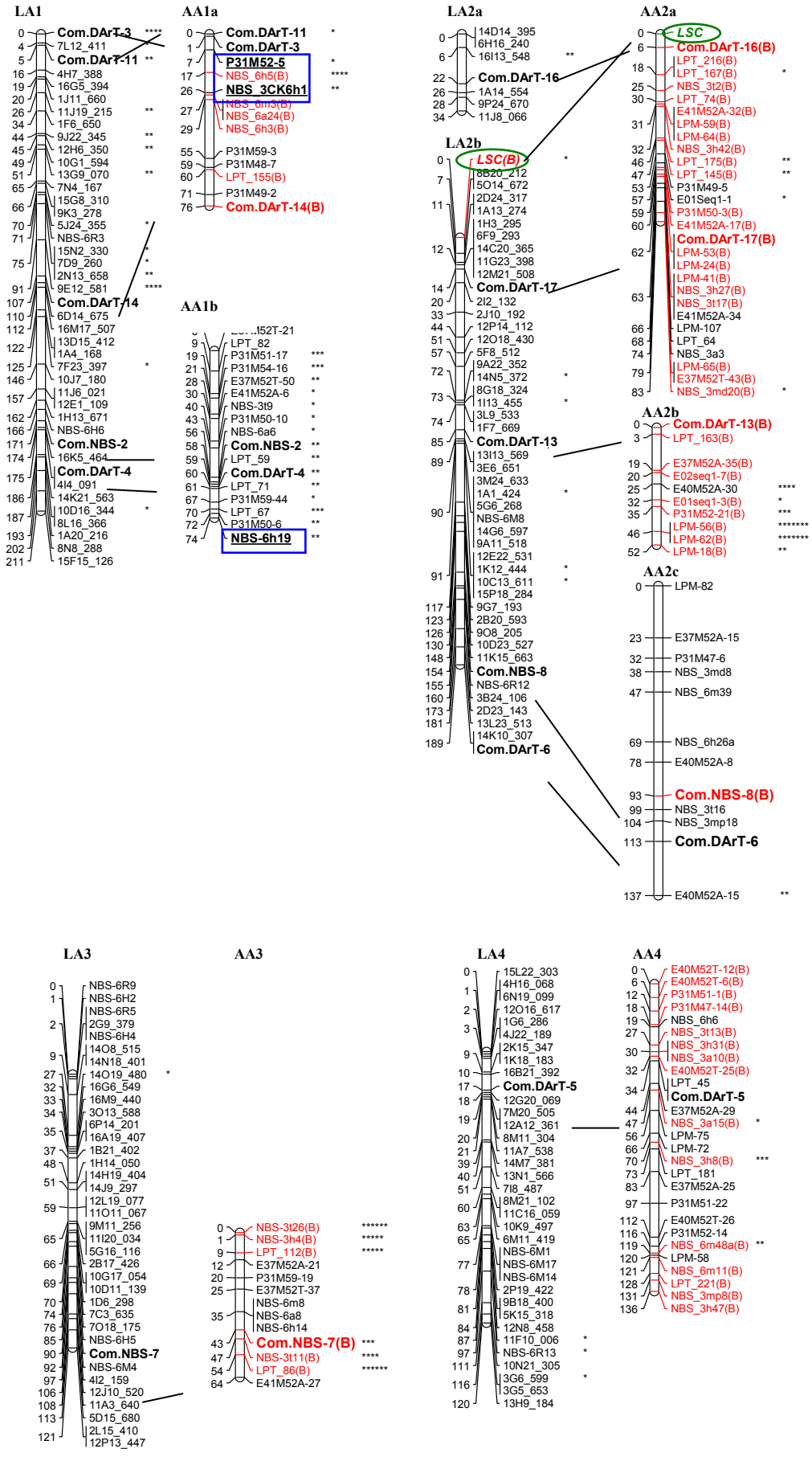
Construction of the genetic linkage maps of the two populations

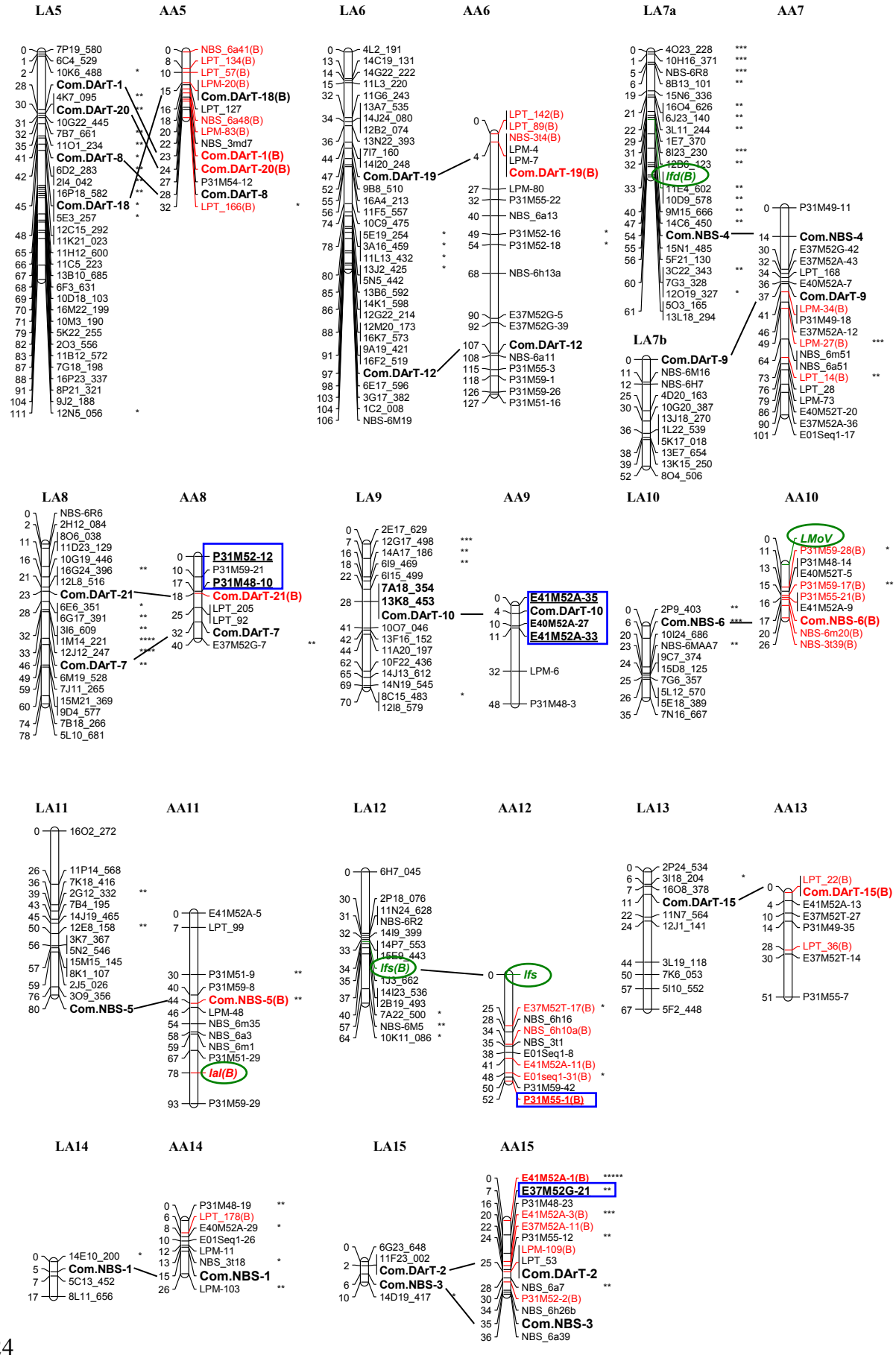
LA population

A total of 34 NBS and 653 DArT markers were available to construct the genetic map. Linkage groups for 'Connecticut King' were constructed with a LOD >4. Even though, 101 DArT markers (15% of total) were removed due to identical scoring, considerable redundancy (*i.e.* close localisation) among DArT markers was recorded. In mapping such redundant markers behave as blocks that hinder the ability of the mapping software to choose the best order of the markers in any linkage group. For that, clustered markers that had similar phase and showed identical scores to each other except for few missing values were considered as duplicates and reduced to one marker. This resulted in the removal of another 120 DArT markers. Finally, a total of 411 molecular markers (31 NBS and 380 DArT markers) and three horticultural traits (stem color, flower direction, and spots) were mapped on 22 linkage groups whereas 50 markers (around 12%) remained ungrouped. The mean Chi-square of the linkage groups ranged between 0.14 and 1.6. The LA map spans 1627 cM with an average marker density around 3.9 cM (Fig. 1). Out of 414 markers, 90 showed a skewed segregation (21.7 %), of which many clustered on linkage group LA 7a.

AA population

Linkage groups of 'Connecticut King' were constructed for AA population. The AA genetic map was extended by adding 155 NBS and 244 DArT markers to 301 previously assessed AFLP markers (183 <nnxnp> and 118 <hkxhk>). For basic mapping step only the markers that segregated in a 1:1 fashion were used. This step was done to avoid any arbitrary linkage that might occur due to the <hkxhk> type markers. Then markers order was fixed, and <hkxhk> markers were added. However, <hkxhk> markers that caused suspect linkage or considerable tension in the map were excluded. In this cross, many markers remained ungrouped (177 markers, around 25% of total number, 131 of <lmxll> type and 46 of <hkxhk> type). In total, 295 markers (134 AFLP, 70 NBS, and 91 DArT markers) and five traits (LMoV, stem color, antherless, flower color, and spots) were mapped with a LOD > 4. This map comprises 24 linkage groups (LGs) of 6 or more markers (32 markers grouped in very small LG with 2 till 5 markers). The mean Chi-square of the linkage groups ranged between 0.5 and 2. The AA genetic maps spans 1517 cM with an average density of a marker each 5 cM. Overall 76 markers out of 300 were skewed (25.5 %) and clustered mainly on LG 1 and 3 (Fig. 1).


LA1
0 Com.DArT-3 ****
4 7L12_411
5 Com.DArT-11 **
16 4H7_388
19 16G5_394
20 1J11_660
26 11J19_215 **
34 1F6_650
44 9J22_345 **
45 12H6_350 **
49 10G1_594
51 13G9_070 **
65 7N4_167
66 15G8_310
9K3_278
70 5J24_355 *
71 NBS-6R3
15N2_330 *
75 7D9_260 *
2N13_658 **
91 9E12_581 ****
107 Com.DArT-14
110 6D14_675
112 16M17_507
122 13D15_412
1A4_168
125 7F23_397 *
146 10J7_180
157 11J6_021
12E1_109
162 1H13_671
166 NBS-6H6
171 Com.NBS-2
174 16K5_464
175 Com.DArT-4
4I4_091
186 14K21_563
187 10D16_344 *
8L16_366
193 1A20_216
202 8N8_288
211 15F15_126
AA1a
0 Com.DArT-11 *
1 Com.DArT-3
7 P31M52-5 *
17 NBS_6h5(B) ****
26 NBS_3CK6h1 **
27 NBS_6m3(B)
NBS_6a24(B)
29 NBS_6h3(B)
55 P31M59-3
59 P31M48-7
60 LPT_155(B)
71 P31M49-2
76 Com.DArT-14(B)
AA1b
9 LPT_82
19 P31M51-17 ***
21 P31M54-16 ***
28 E37M52T-50 **
30 E41M52A-6 *
40 NBS-3t9 *
43 P31M50-10 *
56 NBS-6a6 *
58 Com.NBS-2 **
59 LPT_59 **
60 Com.DArT-4 **
61 LPT_71 **
67 P31M59-44 *
70 LPT_67 ***
72 P31M50-6 **
74 NBS-6h19 **
LA2a
0 14D14_395
6H16_240
6 16I13_548 **
22 Com.DArT-16
26 1A14_554
28 9P24_670
34 11J8_066
LA2b
0 LSC(B) *
7 8B20_212
5O14_672
11 2D24_317
1A13_274
1H3_295
6F9_293
12 14C20_365
11G23_398
12M21_508
14 Com.DArT-17
20 2I2_132
33 2J10_192
44 12P14_112
51 12O18_430
57 5F8_512
9A22_352
72 14N5_372 *
8G18_324
73 1I13_455 *
74 3L9_533
1F7_669
85 Com.DArT-13
89 13I13_569
3E6_651
3M24_633
1A1_424 *
5G6_268
90 NBS-6M8
14G6_597
9A11_518
12E22_531
91 1K12_444 *
10C13_611 *
15P18_284
117 9G7_193
123 2B20_593
126 9O8_205
130 10D23_527
148 11K15_663
154 Com.NBS-8
155 NBS-6R12
160 3B24_106
173 2D23_143
181 13L23_513
14K10_307
189 Com.DArT-6
AA2a
0 LSC
6 Com.DArT-16(B)
LPT_216(B)
18 LPT_167(B) *
NBS_3t2(B)
25 LPT_74(B)
30 E41M52A-32(B)
31 LPM-59(B)
LPM-64(B)
32 NBS_3h42(B)
46 LPT_175(B) **
47 LPT_145(B) **
53 P31M49-5
57 E01Seq1-1 *
59 P31M50-3(B)
60 E41M52A-17(B)
Com.DArT-17(B)
62 LPM-53(B)
LPM-24(B)
LPM-41(B)
63 NBS_3h27(B)
NBS_3t17(B)
E41M52A-34
66 LPM-107
68 LPT_64
74 NBS_3a3
79 LPM-65(B)
E37M52T-43(B)
83 NBS_3md20(B) *
AA2b
0 Com.DArT-13(B)
3 LPT_163(B)
19 E37M52A-35(B)
20 E02seq1-7(B)
25 E40M52A-30 ****
32 E01seq1-3(B) *
35 P31M52-21(B) ***
46 LPM-56(B) *******
LPM-62(B) *******
52 LPM-18(B) **
AA2c
0 LPM-82
23 E37M52A-15
32 P31M47-6
38 NBS_3md8
47 NBS_6m39
69 NBS_6h26a
78 E40M52A-8
93 Com.NBS-8(B)
99 NBS_3t16
104 NBS_3mp18
113 Com.DArT-6
137 E40M52A-15 **
LA3
0 NBS-6R9
1 NBS-6H2
NBS-6R5
2 2G9_379
NBS-6H4
14O8_515
9 14N18_401
27 14O19_480 *
32 16G6_549
33 16M9_440
34 3O13_588
6P14_201
35 16A19_407
37 1B21_402
48 1H14_050
51 14H19_404
14J9_297
59 12L19_077
11O11_067
65 9M11_256
11I20_034
5G16_116
66 2B17_426
69 10G17_054
10D11_139
70 1D6_298
74 7C3_635
76 7O18_175
85 NBS-6H5
90 Com.NBS-7
92 NBS-6M4
97 4I2_159
106 12J10_520
108 11A3_640
113 5D15_680
2L15_410
121 12P13_447
AA3
0 NBS-3t26(B) ******
1 NBS-3h4(B) *****
9 LPT_112(B) *****
12 E37M52A-21
20 P31M59-19
25 E37M52T-37
NBS-6m8
35 NBS-6a8
NBS-6h14
43 Com.NBS-7(B) ***
47 NBS-3t11(B) ****
54 LPT_86(B) ******
64 E41M52A-27
LA4
0 15L22_303
4H16_068
1 6N19_099
2 12O16_617
1G6_286
3 4J22_189
2K15_347
9 1K18_183
10 16B21_392
17 Com.DArT-5
18 12G20_069
7M20_505
19 12A12_361
20 8M11_304
21 11A7_538
39 14M7_381
40 13N1_566
51 7I8_487
8M21_102
60 11C16_059
63 10K9_497
65 6M11_419
NBS-6M1
77 NBS-6M17
NBS-6M14
78 2P19_422
81 9B18_400
5K15_318
84 12N8_458
87 11F10_006 *
97 NBS-6R13 *
111 10N21_305
116 3G6_599 *
3G5_653
120 13H9_184
AA4
0 E40M52T-12(B)
6 E40M52T-6(B)
12 P31M51-1(B)
18 P31M47-14(B)
19 NBS_6h6
27 NBS_3t13(B)
NBS_3h31(B)
30 NBS_3a10(B)
32 E40M52T-25(B)
LPT_45
34 Com.DArT-5
44 E37M52A-29
47 NBS_3a15(B) *
56 LPM-75
66 LPM-72
70 NBS_3h8(B) ***
73 LPT_181
83 E37M52A-25
97 P31M51-22
112 E40M52T-26
116 P31M52-14
119 NBS_6m48a(B) **
120 LPM-58
121 NBS_6m11(B)
128 LPT_221(B)
131 NBS_3mp8(B)
136 NBS_3h47(B)


LA5
AA5
LA6
AA6
LA7a
AA7
LA7b
LA8
AA8
LA9
AA9
LA10
AA10
LA11
AA11
LA12
AA12
LA13
AA13
LA14
AA14
LA15
AA15

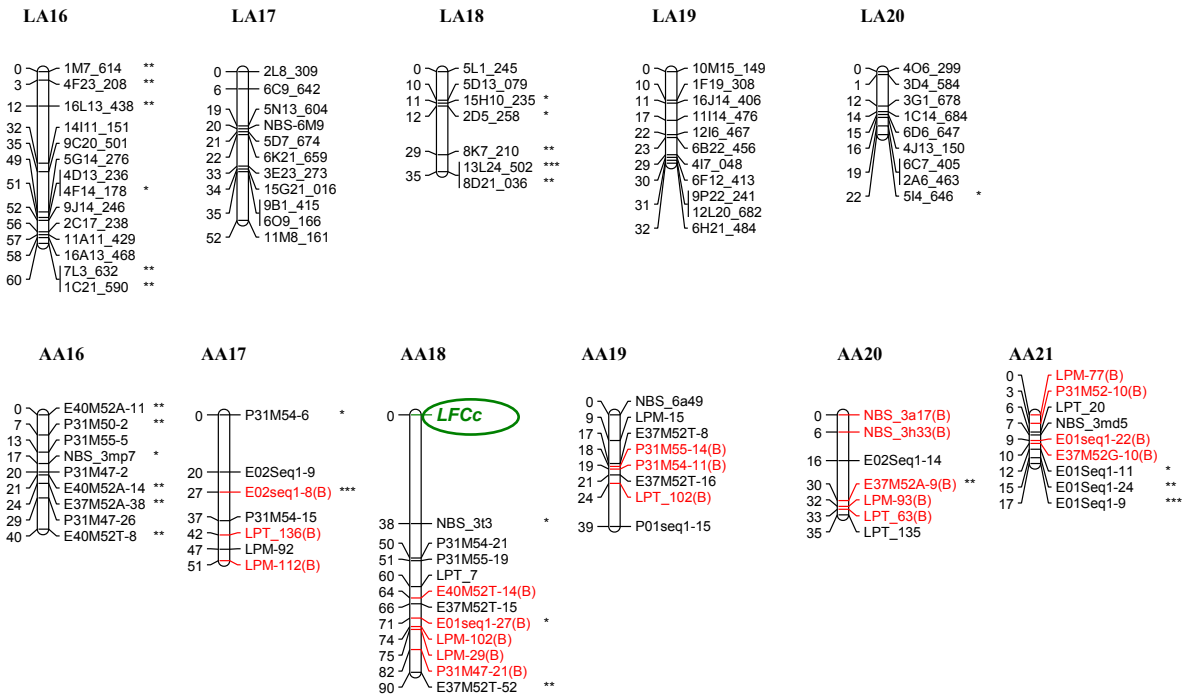


Figure 1. Aligned genetic maps of the LA and AA populations. The * refers to the skewedness of the markers at P-value ranged between 0.05 and 0.0001. All <hkxhk> markers are followed by a (B) and colored in red. All bold markers are common markers (Com.DArT-1 till 21, and Com.NBS-1till 8). The mapped traits (*LFCc*, *lfs*, *LSC*, *lal*, *lfd*, and LMoV) were encircled. The six putative quantattive trait loci for *Fusarium* resistance were placed in rectangles.

Map alignment

Since there is no reference genetic maps with transferable markers in lily that can be used to compare our linkage groups with, we characterize our linkage groups according to their length, using those of the LA linkage groups as a reference. The identified common markers were used to find corresponding linkage groups in AA populations and to name them accordingly, *e.g.* linkage groups from the two populations containing the markers Com.DArT-3, Com.DArT-11, and Com.DArT-14 were named as LA1 and AA1 in LA and AA populations, respectively. The first fifteen linkage groups (1 untill 15) of both populations have common markers and are thus aligned. The common markers proved to be very useful in aligning separate small linkage groups of one population to a single larger linkage group of the other population. For instance, five common markers (Com.DArT-3, Com.DArT-11, Com.DArT-14, Com.DArT-4, and Com.NBS-2) were mapped on LA1. As they were mapped in two separate AA linkage groups, this indicates that these two AA linkage groups belong to the same chromosome. Consequently they were named as AA1a and AA1b (Fig. 1). The same applies for AA2a, AA2b, and AA2c assigned to LA2a and LA2b. As a result, the number of linkage groups was reduced from 22 and 24 to 20 and 21 in the LA and AA populations, respectively.

Analysis and mapping of phenotypic traits

Flower color in the Asiatic population was scored visually (orange or yellow). Remarkably, the color of the basal part is different in some genotypes (28 genotypes) from the upper part; however, a very high correlation was detected (98%) between these two parts. For that, just the color of the upper part of the flower was scored. The trait segregated 1:1 (Table 1) and was mapped on the top of AA18, 38 cM from the closed marker (Fig. 1).

Absence/presence of spots in LA population segregated in a 3:1 ratio and in a 1:1 ratio in AA population (Table 1) which indicates that the presence of spots on lily flowers is controlled by a recessive gene that we propose to denote as *lfs* (*Lilium* Flower Spot). The two spotless parents 'Connecticut King' and 'White Fox' are heterozygous whereas 'Orlito' is homozygous for the recessive allele having spots. The *lfs* locus was mapped on LA12 as a marker very close to two DArT markers (1 cM far from 1J3_662 and 15E9_443); however, on AA12 the closest marker was 25cM far (E37M52T-17). Even though one major gene controls the presence/absence of spots, the variation in spots number indicates the presence of additional genes with minor effects controlling spots number. The number of spots varied between 1 and 50 in LA and between 1 and 44 in AA populations and was mapped as a continuous trait using MapQTL. In both populations, spots number showed a very strong QTL exactly at the same position of the mapped *lfs* locus (LG12). In LA population, the QTL (LOD 6.9, threshold of 2.75) explained around 30% of the phenotypic variation whereas in AA population, the QTL (LOD = 8.4; threshold of 4.6) explained 33% of the variation. In both populations we could not detect any other QTLs that might have minor effects on the number of spots. Next, all the individuals that have no spots in AA population (51 individuals) were excluded and the data from the 46 individuals left were transformed (^{10}Log). This step was done to identify minor QTLs which might be masked by the effect of the major QTL on LG12. However, once more only the region of the *lfs* gene was identified with a LOD of 4.9 (threshold was 4.4) and it explained 36.5% of the variation in spots numbers, while the closest marker (E37M52T-17 at 25 cM) did not show any correlation with spot numbers.

In LA population stem color segregated 3:1 (green 77: 15 dark green, χ^2= 3.7, *P*=0.05), while in AA population the segregation was 1:1 (53 green: 46 dark green, χ^2= 0.5; Table 1). This result indicates that the variation in stem color is based on a single gene for which green is dominant. We propose to denote this gene as *LSC* (*Lilium* Stem Color). Both green parents 'Connecticut King' and 'White Fox' must be heterozygous. The dark green parent 'Orlito' (and 'Pirate') is homozygous recessive. *LSC* could be mapped as a <hkxhk> marker on LA2b and as <lmxll> marker on AA2a at 6 to 7 cM from the nearest marker in both maps.

The antherless phenotype (*lal*, *Lilium* antherless mutation) in AA population segregated in a 3:1 ratio (77 for the presence of the anther: 20 for the absence of the anther, χ^2=1.01), which indicates that it is a single gene based trait. Since both parents have anthers and since their progeny segregates, both parents should be heterozygous and the locus regulating the *lal* phenotype

should be recessive. This trait was mapped as a <hkxhk> marker type on AA11 with 11 cM from the nearest marker (Fig. 1). The same holds true for the gene influencing flower direction *lfd* (*Lilium* Flower direction) in LA population (63 out-faced: 27 up-faced, χ^2=1.2), a recessive allele that results in up-faced flowers. This was mapped on LA7a (co-localised with two DArT markers).

Table 3: The QTLs of *Fusarium* resistance as a result of four years of disease test. -, *, **, ***, ****, ***** and ****** refer to non-significant, significant at *P*= 0.1, 0.05, 0.01, 0.005, 0.001, and 0.0005 respectively as a result of Kruskal-Wallis test.

Putative QTL	Linkage group	Position on linkage group	1992	1993	1994	1999
QTL1	AA8	P31M52-12, P31M48-10	******	****	**	*****
QTL2	AA9	E41M52A-35, E41M52A-33	****	*	-	****
QTL3	AA12	P31M55-1	***	****	*	**
QTL4	AA15	E37M52G-21	**	**	**	-
QTL5	AA1b	NBS-6h19	-	****	**	-
QTL6	AA1a	NBS-3CK6h1, P31M52-5	***	**	-	**

LMoV resistance was considered as a monogenetic trait since it segregated as 1:1 and mapped as a marker on AA10. As for *Fusarium* resistance, six putative QTLs were identified in the Kruskal-Wallis test. The first four, QTLs mapped on linkage groups AA8, AA9, AA12, and AA15 respectively (Fig. 1) were similar to the ones identified by Van Heusden et al. (2002). In addition, two new putative QTLs (QTL 5 and 6), mapped on AA1b and AA1a respectively, were identified in the current study. Not all QTLs were detected in each of the four years; also they varied in their significance level (Table 3). For example, in 1993 and 1994 QTL5 was identified as a significant QTL while it was not detected in 1992 and 1999. On contrast, QTL2 was absent in 1993 and 1994 and was identified in 1992 and 1999. Nevertheless, QTL1 seems a strong and reliable QTL, since it was detected in the four years and showed to be the most significant. By interval mapping followed by rMQM, only QTL1 of AA8 (LOD=6) exceeded the threshold LOD score (4.3) in 1992 explaining about 25% of the phenotypic variation.

Discussion

Map construction

In current study, we were able to generate for the first time NBS markers and DArT libraries in lily. This allowed us to construct genetic maps for a species with one of the largest genomes in the animal and plant kingdoms. NBS markers are well distributed over the linkage groups. Some clustering is noticed (*e.g.* see LA3), which is similar to the resistance gene analogues clustering found in many other species such as apple and rapeseed (Calenge et al., 2005; Tanhuanpää, 2004). DArT markers have been useful for constructing linkage maps in barley (Wenzl et al.,

2004), cassava (Xia et al., 2005), *Arabidopsis* (Wittenberg et al., 2005), pigeon pea (Yang et al. 2006), wheat (Akbari et al., 2006; Mantovani et al., 2008), and *Sorghum bicolor* (Mace et al., 2008). In lily, this has a much larger genome than any of the previously mentioned crops, DArT technology also proved to be a highly efficient method for map construction. Redundancy of the DArT markers was recorded in LA population which was also found in *Arabidopsis* (Wittenberg et al., 2005). This can be explained by repeated representation of the same *Pst*I-*Pst*I fragment on the array (*e.g.* 7A18_354, 13K8_453, and Com.DArT-10 on LA9), which may be caused by sampling high numbers of clones from a relatively limited pool due to a bias in cloning efficiency and unequal amplification efficiency of DNA fragments. Likewise, clustering of markers due the presence of repeat regions with *Pst*I sites can also lead to redundancy. Advantages of DArT marker are that many markers can be produced in a single effort at low cost, scoring is automated, and the markers are highly reliable and transferable. Moreover, it allows to concentrate on gene rich regions of the genome (by using a methylation sensitive restriction enzyme like *Pst*I), a characteristic which is of high advantage for genomes in which a high percentage of the genome is repetitive, like in *Lilium.* Similar to AFLP and NBS markers, DArT markers also do not require prior sequence information. However DArT marker technology does require serious initial investment in array development and it is technically more demanding in application and in laboratory equipment, due to which outsourcing may be more frequently needed. The various marker systems (AFLP, NBS, and DArT) proved equally efficient in the proportion of initial markers that could be mapped. However, the DArT system outcompeted the other systems in ease of data generation and reliability of scores, the others requiring manual scoring which on itself was time demanding and also showed to be prone to scoring and administrative errors.

One additional essential advantage of DArT is the ease of marker conversion into a single locus marker for downstream MAB applications. The first step of converting any marker of interest (*i.e.* linked to important trait) is often to reproduce this marker. This can be challenging in molecular marker systems such as AFLP and NBS profiling due to possible co-migrating bands, and be even more difficult in RAPD markers that are only reproducible under very strictly controlled conditions within one and the same laboratory (Jones et al., 1997). As DArT markers are based on cloned fragments, the sequencing of these markers can be performed directly without the need to reproduce a specific fragment. Furthermore, the larger length of DArT fragments (average in our study 1 kb), compared to AFLP and NBS bands, enhances the chance to find polymorphisms. Moreover the hybridization step is helpful in selection against highly repetitive sequences, which supports ease of conversion as well as applicability of the resulting markers. Indeed, thus far highly repetitive sequences hampered with conversion of AFLP and NBS bands into single locus markers from lily so far (data not shown).
For mapping, only 1:1 markers were used in LA population, while both 1:1 and 3:1 marker types were used in AA population. The latter was because in AA backcross population more than half of the loci will segregate as <hkxhk>. If <hkxhk> markers would be discarded a large part of the 'Connecticut King' genome would remain uncovered (Fig. 1).

A considerable number of AFLP, NBS, and DArT markers used in this study deviated ($P<0.05$, Chi-square test) from the 1:1 and 3:1 ratio expected in these two interspecific populations (21.7% and 25.5% in LA and AA populations, respectively). A similar ratio (24-29%) of distorted markers has been identified by (Abe et al., 2002). Distorted marker segregation is a common feature in most interspecific crosses as has been reported in crops such as, maize (Sibov et al., 2003), barley (Konishi et al., 1992), sugar beet (Pillen et al., 1993), and coffee (Ky et al. 2000). Some of the distorted markers are clustered in linkage groups (*e.g.* on AA3) which might be due to the presence of segregation distortion loci (SDLs) on these particular linkage group. Different reasons can lead to this phenomenon in interspecific lily hybrids such as: the presence of lethal genes, parental reproductive differences, and chromosomal rearrangement (Blanco et al., 1998; Fauré et al., 1993; Foolad et al., 1995). The selective effect of SDLs determines the direction of skewed markers and extent of skewedness. The closer the linkage between markers and SDLs, the larger the Chi-square value (Xian-Liang et al., 2006). Such pattern can be seen in a number of instances in the present study. On the other hand, a number of isolated markers distributed over different linkage groups show a deviation in their segregation pattern (LA2a and AA2a). Segregation distortion at single isolated marker locus can be related to errors in marker genotyping or a point mutation at the binding site for the DNA marker (Sibov et al., 2003; Smith et al., 1997).

The number of linkage groups in the two populations exceeded the haploid chromosome number (x=12). This is similar to previous studies in lily in which 24 and 26 linkage groups were constructed (Abe et al., 2002). In some other crops, also, the number of linkage groups exceeded chromosome counts such as *Catharanthus roseus* (2n= 2x=16) with 14 linkage groups (Gupta et al., 2007). Lily has a very large genome compared with other plant genomes, moreover, lily has an exceptional high chiasmata frequency of 54.8 per cell (Stack et al., 1989) compared with other plant species such as 19 in onion (Albini and Jones, 1990) and 28 in *Zea mays* (Gillies, 1983). This phenomenon has also been recorded in chicken (x=39) where around 59–64 chiasmata per cell were found (Rodionov et al., 2002) and the mapping efforts resulted in 50 linkage groups (Groenen et al., 2000). This might be explained by hot spots on some chromosomes that result in the division of such a chromosome in several linkage groups. For lily, chromosome nine shows a high recombination frequency (Khan et al., 2009) which might indicates the precence of a recombination hotspot. The high chiasmata frequency, moreover, might also result in large linkage groups. As consequence, several of the small linkage groups (less than 100 cM) are expected to belong together.

The expected length of a genetic map can be calculated using the hypothesis that one chiasma correspond to 50 cM (Rodionov et al., 2002). An average of 54.8 chiasmata was determined per complete set of diplotene bivalents in *L. longiflorum* (Stack et al., 1989) which approximates to a calculated total genome length of 2740 cM. Therefore, the genetic map of LA population (1627 cM) covers around 60% of lily genome compared with 55% (1517 cM) for AA population. This

suggests that more markers are needed to reach a high density genetic map that covers the whole lily genome.

With long chromosomes, under-representation of markers in particular chromosomal segments may easily occur. This might be particularly true for LA population, where the use of methylation sensitive *Pst*I restriction endonuclease for genome complexity reduction in DArT marker production, will likely lead to under-representation of markers in methylated parts of the genome. NBS profiling was also applied in this population, however, due to some technical problem with DNA quality, relatively few markers (34) were produced that are unlikely to cover the methylated part of the genome in a substantial way. For that, maybe producing DArT markers using non-methylation sensitive restriction enzyme which tags the whole genome can help in better coverage of the genome, although this may need the use of a third restriction enzyme to accomplish the genome complexity reduction needed. In AA population, where also *Eco*RI/*Mse*I AFLP markers have been used (along with *Pst*I/*Mse*I), better coverage can be expected. The regions of low marker density (gaps) leading to splits into different LGs may here be associated with either genomic regions that are identical by descent, genomic regions that have very limited genetic variability in the initial diversity representation, or similarly has hot spots with a high recombination frequency.

Common markers between the two crosses were helpful in aligning and linking linkage groups to each other. Thus the number of linkage groups reduced from 22 and 24 to 20 and 21 in LA and AA populations, respectively. Unfortunately, the number of common markers was insufficient to construct a consensus map of 'Connecticut King' from these two populations.

Traits mapping

Very few studies describe the mapping of genes that control economic important traits in ornamental plants, such as leaf size, flowering time, and flower size. These studies are maily in roses (Dugo et al., 2005), and petunia (Galliot et al., 2006). In this study, we mapped and identified markers associated with genes that control several morphological traits in *Lilium:* flower color, flower spots, stem color, antherless phenotype, and flower direction in *Lilium*, as well as resistance to Lily mottle virus and *Fusarium*.

Flower-anthocyanin and tepal-carotenoid pigmentation traits were each found to be controlled by a major QTL in apple, rose, and *Rhododendron* (Cheng et al., 1996; Debener and Mattiesch, 1999; Dunemann et al., 1999). Similar results were obtained in lily where each of both traits was found to be controlled by a major QTL that were mapped on the genetic map of cv. ‘Montreux’ (Abe et al., 2002; Nakano et al., 2005). The pink color of cv. ‘Montreux’ is related to high levels of anthocyanin and very low levels of carotenoid. Therefore, the two QTL’s identified in ‘Montreux’ were proposed to be for anthocyanin pigmentation and for suppression of carotenoid (Abe et al., 2002; Nakano et al., 2005). However, both the yellow and orange colors of Asiatic cultivars are related to carotenoid pigments, which is highly expressed in ‘Connecticut King’

(Yamagishi et al., 2010) whereas a little or no anthocyanin pigment was detected in this cultivar (Abe et al. 2002). As a consequence, the *LFCc* (*Lilium* Flower color carotene) locus that was mapped on AA18 and explains the variation in flower color in our population between yellow and orange, might be a carotenoid gene or a transcription factor that regulate the expression of carotenoid pigment in this population (Koes et al., 2005). The flower color locus mapped at the top of AA18, 38 cM from the nearest marker. Mapping at the extremes of a LG at considerable distance of the most distant marker can indicate mistakes in phenotype scoring but this was not likely here because the flower color was checked back and it is an easily scorable trait. Thus the large distance to the nearest marker meight be due to a local under-representation of markers or to a presence of a hot spot in this region both of which would give the same result.

A single locus was identified in both populations (LG12) that represents the spotless allele in 'Connecticut King'. Abe et al. (2002) identified two QTLs when mapping the number of spots as a continuous trait. The continuous distribution of flower spots indicates that several genes regulate this trait. However, since the presence of spots in this study segregated as 1:1, it means that there is a single major gene that controls the formation of spots whereas the number of spots may be controlled by other genes with a minor effect. In this study, only one single locus was identified for both the presence/absence of spots and for the variation in spot number. This might due to a single gene controlling both traits, or due to the involvement of two or more closely linked genes. However, the possibility also remains that other genes are involved which may be located on regions not yet covered by markers.

Stem color was mapped as a marker at the top of LA2b and AA2a in both populations. However, comparing both populations for the order of the common markers and the trait shows an inconsistency in the location of this trait. In the AA2a, stem color is linked to Com.DArT-16, while in LA2b the same common marker is not clearly associated with this trait. This inconsistency may be due to the high involvement of <hkxhk> type markers in AA population and the <hkxhk> type segregation of the trait in LA population. As for this segregation type only 25% of the data are informative from a genetic point of view, marker positions are not very accurate.

Interestingly, we were able to map the antherless phenotype. It was reported that Asiatic lily hybrids can have the antherless mutation which leads to antherless stamen. Several factors have been found to regulate the expression and reversal of the antherless phenotype such as high temperature in lily (32ºC/25ºC day/night) (Sato and Miyoshi, 2007), and gibberellins in tomato (Phatak et al., 1966) which may indicate an epigenetic mutation. Since all individuals received exactly the same treatments in all stages in our study, the mapped trait (AA11) here is related to a mutated gene which is responsible for anther formation in lily. A similar mutation (*afo-1*) was found in *Arabidopsis* by insertion screening of a Ds transposable element (Kumaran et al., 1999). A significant correlation (r^2= 0.39) was identified between antherless phenotype and flower spots formation in lily by (Straathof et al., 1996). Most antherless individuals in our study do not have

spots (in few cases one or two spots are present). However, in AA map the two single gene based traits (antherless and flower spots) were mapped on different linkage groups (AA11 and AA12) suggesting an independent segregation from each other. This might be due to the fact that our maps are not completely saturated. The correlation between the two traits is not very strong (r^2= 0.39), which might means that they are not localized close to each other on the chromosome. If this chromosome is represented by two LGs due to a lack of markers that can bridge between them, then AA11 and AA12 might be representing different regions of the same chromosome.

Flower direction is an economically important trait in the Longiflorum group, since the common out or down-facing phenotype leads to flower damage and quality losses in packaging and higher transport costs. The flowers in LA population show a range of angles varying between up-facing till out-facing. Grouping this trait into two groups (up and out-face) helped in mapping it on LA7a. However more accurate measurement of this trait is needed to figure out the continuous segregation of this trait.

For breeding, the availability of markers for recessive ornamental traits such as spots, antherless, flower direction is very useful. Such markers allow the identification of suitable breeding parents so that expression of the recessive trait can be either enhanced or repressed.

Disease resistance (re-)mapping in the AA map slightly changed the results of the mapping presented by Van Heusden et al. (2002). For the LMoV a few NBS and DArT markers were added to the linkage group. For the *Fusarium* resistance, apart from additional NBS and/or DArT markers and slight changes in marker order in the four previously identified QTL regions, two additional putative QTLs were identified. Interestingly, significant association was found between resistance and some of the common markers. For example: Com.DArT-10 that mapped on AA9 is significantly associated with the resistance QTL2 whereas in LA map the marker is closely linked to two other DArT markers (7A18_354 and 13K8_453). This may help us to identify the position of QTLs related to *Fusarium* resistance in LA population even before actual mapping of this trait there. Unfortunately, only QTL1 that explained a considerable proportion of the phenotypic variation was confirmed by interval mapping test in one year (1992). Also with the Kruskall-Wallis approach, not all the putative QTLs were significantly linked to resistance in the four years. Thus, it is an essential need to run disease test in a number of consecutive years, and to control all the possible variations that might play a role in the quality of the test to obtain accurate QTL mapping results.

Chapter 3

SNP Markers Retrieval for a Non-Model Species: A Practical Approach

Arwa Shahin[1], Thomas van Gurp[3], Sander A. Peters[2], Richard G.F. Visser[1], Jaap M. van Tuyl[1], Paul Arens[1]

[1] Wageningen UR Plant Breeding, Wageningen University and Research Centre, P.O. Box 386, 6700 AJ, Wageningen, the Netherlands

[2] BU Bioscience Plant Research international, Wageningen University and Research Centre, P.O. Box 16, 6700 AJ, Wageningen, the Netherlands

[3] Netherlands Institute of Ecology (NIOO-KNAW), Department of Terrestrial Ecology, P.O. Box 50, 6700 AB Wageningen, the Netherlands

This chapter has been published with modification in BMC Research notes (2012) 5:79

Abstract

SNP (Single Nucleotide Polymorphism) markers are rapidly becoming the markers of choice for applications in breeding because of next generation sequencing (NGS) technology developments. For SNP development by NGS technologies, correct assembly of the huge amounts of sequence data generated is essential. Little is known about assembler's performance, especially when dealing with highly heterogeneous species that show a high genome complexity and what the possible consequences are of differences in assemblies on SNP retrieval. This study tested two assemblers (CAP3 and CLC) on 454 data from four lily genotypes and compared results with respect to SNP retrieval. CAP3 assembly resulted in higher numbers of contigs, lower numbers of reads per contig, and shorter average read lengths compared to CLC. Blast comparisons showed that CAP3 contigs were highly redundant. Contrastingly, CLC in rare cases combined paralogs in one contig. Redundant and chimeric contigs may lead to erroneous SNPs. Filtering for redundancy can be done by blasting selected SNP markers to the contigs and discarding all the SNP markers that show more than one blast hit. Results on chimeric contigs showed that only four out of 2,421 SNP markers were selected from chimeric contigs. In practice, CLC performs better in assembling highly heterogeneous genome sequences compared to CAP3, and consequently SNP retrieval is more efficient. Additionally a simple flow scheme is suggested for SNP marker retrieval that can be valid for all non-model species.

Introduction

In last few years, the development of next-generation sequencing technologies that have the capacity to generate millions of short reads in a single run, has led to a revolution in sequencing applications. The NGS technologies not only boosted re-sequencing and allele mining studies in model species, but are also very useful for the development of SNP markers in species with no or hardly any genetic resources.

SNP development using NGS technologies essentially has become cheaper and faster but also generated requirements like the need for genome complexity reduction, assembly of sequences, and SNP identification in high throughput. The latter two steps are still considered challenging. Currently many different assemblers are available, but few studies discussed the performance of different assemblers in relation to assembly quality and the influence of genome complexity and heterogeneity on the quality of the assembly. Assembly quality is generally assessed by: the lengths of the contigs (mean, minimum and maximum lengths, or N50 according to the assembler), and the accuracy or correctness of the assembly (how well the contigs can be mapped to the reference genome) (Paszkiewicz and Studholme, 2010). Two different assemblers (Newbler and MIRA) were compared on an insect sequence dataset using public Sanger EST data and 454 transcriptome data (Papanicolaou et al., 2009). Another study compared six assemblers (CAP3, MIRA, Newbler 2.3, Newbler 2.5, SeqMan, and CLC) in reference to the number and

length of contigs, speed of assembly and assembly redundancy in *de novo* assembly of a nematode (Kumar and Blaxter, 2010). The quality of the contigs was checked by aligning the contigs to four reference sequence sets (ESTs, proteome, gene families, and protein data from databases). Similarly, the performance of six aligners (BLAT, SSAHA2, Bowtie, SeqMap, MAQ, and CLC) were compared using *in silico* generated transcripts from four model organisms (human, *Arabidopsis*, *Drosophila* and yeast) that were mapped to the transcriptome or the complete genome from sequence databases (Palmieri and Schlötterer, 2009). Results showed that with increasing sequence read length mapping was more accurate, while with increasing genome heterozygosity more reads were incorrectly mapped. Recently, a comparison in which eight short reads assemblers were evaluated against two types of simulated short reads datasets (allowing 0.1 % error rate) derived from four different genomes (nematode, yeast, bacteria, and virus), was published (Zhang et al., 2011). The assemblers' performance information about computational time, memory cost, assembly accuracy and completeness and size distribution of assembled contigs where studied (by mapping to reference genomes) (Zhang et al., 2011). All these studies used relatively small sized genomes, and often inbred organisms and studied assembly accuracy in general parameters and by mapping to reference genomes or Sanger sequencing data (Kumar and Blaxter, 2010; Palmieri and Schlötterer, 2009; Papanicolaou et al., 2009; Zhang et al., 2011). Additionally these studies showed that there is currently no commonly accepted and standardized method for performance evaluation of assemblers, none of these studies checked the assembly quality concerning SNP markers retrieval, and no clear guidance for assembler selection was defined. Because we are involved in ornamental breeding where, in general, crops are outcrossing and highly heterogeneous without reference sequences, our goal was to study the effects of two different assemblers on assembly performance and SNP retrieval in heterogeneous outcrossing species by using our model crop lily as an example. Running such a study, conformation of assembly quality by mapping to a reference genome would be optimal. However, species with reference genomes do not represent the same level of heterogeneity and genome complexity as is found in most outbreeding non-model species. In our study, we analyzed a highly divergent sequence dataset of the non-model species lily that allows us to investigate a real case study and develop a flow scheme that can be followed in SNP marker development studies for similar non model species.

When working with *Lilium,* which has an assumed high level of diversity, a large genome size of 36 Gb with an accompanying high genome complexity and a lack of genetic resources, assembly is an important step in SNP retrieval. Since clear criteria on choosing an assembler are lacking, in our study we focused on two widely used assemblers (CAP3 and CLC) which represent the two different approaches which are used in assemblers. CAP3 is selected since it uses the overlap algorithm for assembly and was successfully used to assemble EST genebank data in heterozygous species such as *Zea mays* (Emrich et al., 2004) and potato (Anithakumari et al., 2010; Tang et al., 2006). Recently it was used to assemble apricot (*Prunus armeniaca* L.), castor bean, mulberry (*Morus sp*.), Pigeonpea (*Cajanus cajan* L.), rice and grape (Dubey et al., 2011; Franssen et al., 2011; Gulyani and Khurana, 2011; Rivarola et al., 2011; Sakai et al., 2011; Tillett

et al., 2011; Vera Ruiz et al., 2011). Furthermore, CAP3 is implemented in the QualitySNP pipeline (Tang et al., 2006) which is a pipeline to identify SNPs and was used in SNP mining studies (Anithakumari et al., 2010; Singhal et al., 2011). CLC assembler is selected since it uses the de Bruijn algorithm, it was used in several comparison studies and showed to produce a good quality assembly (Kumar and Blaxter, 2010; Palmieri and Schlötterer, 2009). It is a user friendly assembler since it is not a command line programming software and it has a complete package (cleaning, trimming, clonality removal, SNP and InDels counting, and assembly, in addition to a very advanced visualization technique of the assemblies) which make it a very appealing software to be used. Moreover, CLC assembler supports both short read and long read assembly, and also supports *de novo* assembly of paired end data. Moreover, CLC was used because it was indicated to perform better in mapping of artificial datasets with increased heterogeneity (Palmieri and Schlötterer, 2009). Additionally, recent papers on the performance of assemblers used both assemblers (Bräutigam et al., 2011; Kumar and Blaxter, 2010), which indicates the importance and usability of both assemblers.

In this study CAP3 and CLC were used for *de novo* assembly of 454-transcriptome reads derived from *Lilium*. The goals of this study were: 1) comparing the performance of CAP3 and CLC by running *de novo* assembly, 2) show the influence of the assembler on the reliable detection of alleles and SNPs, and 3) suggesting a simple flow scheme to generate reliable SNP markers out of such heterozygous species.

Materials and Methods

Plant materials

Four lily genotypes that represent the four main hybrid groups of the genus *Lilium* were used for sequencing: cv. 'Star Gazer' (Oriental), breeding line 'Trumpet 061099' (Trumpet), cv. 'White Fox' (*Longiflorum*), and cv. 'Connecticut King' (Asiatic). Young leaves (500 mg) were collected and kept at -80°C upon RNA isolation.

RNA isolation and cDNA library preparation

Using the Trizol protocol (Invitrogen. Carlsbad, CA, USA), the RNA of the four genotypes was isolated and subsequently purified using the RNeasy MinElute kit (Qiagen, Hilden, Germany).
RNA library processing *i.e.* cDNA synthesis, normalization of the cDNA and adaptor ligation for GS FLX Titanium sequencing, was performed by Vertis Biotechnologie AG (Freising, Germany). In short, 45 ug of total RNA of each of the four samples was treated with DNase and then primed with 6 nucleotide random primers for first strand cDNA synthesis. Next, 454 adapters A and B with a unique 6 nucleotides barcode for each cultivar were ligated to the 5' and 3' ends of the cDNAs. These cDNAs were subjected to two steps of PCR: one before the normalization step (around 18 cycles) and one after it (around 8 cycles) using a proof reading enzyme. Normalization was carried out by one cycle of denaturation and re-association of the

cDNAs and subsequent column purification. For Titanium sequencing the cDNAs in the size range of 500 – 600 bp were eluted from preparative agarose gels.

454 sequencing procedures

The four cDNA libraries were mixed in equal concentrations and sequenced on a Life Sciences GS-FLX Titanium according to standard procedures (454 Life Sciences) at Wageningen UR Greenomics (Wageningen, the Netherlands). Raw sequence data are available at ENA-SRA (European Nucleotide Archive-Sequence Read Archive) with the accession number ERP001106.

Assembly

Raw unprocessed sequences were cleaned before assembly using both the reads and the accompanying sequence quality information (SFF files). Trimming was done by removing: 5' and 3' adapters sequences, low quality bases (limit 0.05), ambiguous nucleotides (maximum 2 nucleotides allowed), terminal nucleotides (one nucleotide from the 5' end and 15 nucleotides from the 3' end), and removal of all reads that have less than 100 and more than 800 nucleotides. Next, all the duplicated reads, *i.e.* reads that have the same first 6 nucleotides and exactly the same sequence (>98% similarity), were excluded (clonality) using CD-HIT (Li and Godzik, 2006). After trimming and removing clonality, all the reads were submitted to the standard CAP3 (Huang and Madan, 1999) using the default parameters (threshold identity cutoff 95% over 100 bp) and CLC Genomics Workbench software (CLC bio, Denmark, http://www.clcbio.com/). The *de novo* assembly using CLC was done using the following parameters: conflict resolution (vote), similarity 95% 100 bp over read length and alignment mode (global, do not allow InDels). Through this study few terms will be used frequently such as:

- Assembler's performance: refer to the number of contigs with average contig's length, the number of singletons and assembly redundancy.
- Assembly redundancy: when the assembler tend to separate sequence related to the same locus over different contigs.

SNP detection

All the contigs resulting from CAP3 and CLC were submitted to an updated version of QualitySNP (Tang et al., 2006) to detect reliable single nucleotide variants within each genotype (between the alleles in one genotype, intra SNPs) and between the four genotypes (between the alleles of the four genotypes, inter SNP). SNPs were chosen using the QualitySNP program based on the following criteria: high quality sequence, not within or adjacent to a homopolymeric tract, at least 2 reads of each allele, 50 bp of flanking sequence on each side free of other SNPs and InDels (criteria needed by Illumina Golden Gate platform for SNPs genotyping). Any SNP fitting these criteria is considered and referred to as 'reliable SNP marker', reliable SNP markers are referred as 'high quality' if they are uniquely present in the genome. For the latter, the SNP with 50bp sequence on either side is compared against all contigs of the same assembler using

BLASTN with Expectation value 1E -20. Only SNPs mapped uniquely to the contig from which they were selected (*i.e.* high quality SNPs) will be retained for marker analysis.

Results and Discussion

1. Pre-processing step:

In this study, we generated a large number of genes from the genus *Lilium*. In total, 1,282,735 reads with an average length of 340 bp were derived using 454 pyro-sequencing. The lowest number of reads was obtained from 'Connecticut King' (139,480) reads and the highest from the Trumpet genotype (442,476 reads). From 'White Fox' and 'Star Gazer' 326,539 and 374,240 reads were obtained respectively. This difference in the number of reads might be related to the quality of RNA that was used for each genotype and variations in the initial amount of cDNA that was used of each sample for sequencing.

Cleaning the data showed that 85,719 reads (6.7%) were discarded either because of poor quality, being too short (less than 100 bp), being too long (over 800 bp), or missing the barcode sequence. Around 1,191,938 reads with an average length of 283 bp (after trimming) were kept for further analysis.

Next, all the duplicated reads were removed. The presence of duplicated reads affects the reliability of a SNP call. In sequence data analyses for SNP retrieval, reads are assumed to be from independently derived DNA fragments. Any polymorphism event present independently twice, will be considered as reliable whereas polymorphisms found independently only once could also be due to possible mistakes in cDNA synthesis and PCR steps. Duplicated reads with PCR mistakes still present in sequencing data could result in the selection of these mistakes as SNP and therefore should be avoided. The number of initial transcripts and the effects of differential amplification in the preparation of the sequencing libraries determine the final library output quality (goal is the presence of a variety of transcripts as wide as possible), and thereby affects the percentage of duplicated reads. The more diverse a library is the less duplicated reads. All the clonal reads were excluded (412,826 reads, 35%) and only the longest of the clonal reads were retained leaving a total of 779,112 reads (220,716,355 bp) for the assembly step. Similar results on clonal reads were detected in previous studies in which 11% to 35% of the sequences were reported as potential artificial replicates (*e.g.* (Gomez-Alvarez et al., 2009)). Gomez-Alvarez et al (2009) suggested that this phenomenon could be explained by the binding of amplified DNA fragments generated in the emulsion PCR step of the 454 pyro-sequencing to empty beads. However, clonality of reads is not limited to a specific mechanism since it was recorded in GS20, GS-FLX and GS-FLX Titanium systems (Gomez-Alvarez et al., 2009) as well as in Illumina's Solexa (Kozarewa et al., 2009) which indicates the possibility of another explanation. The relative high clonality found with different sequencing technologies could be related to the cDNA library preparation in which often PCR steps are used to generate sufficient quantities of cDNA for sequencing (Kozarewa et al., 2009). In particular, the second PCR after

the normalization step (using the primers adapters of the A and B adapters) may increase the number of duplicated reads. In our case, we could detect duplicated reads since no shearing of fragments was applied but instead fragments were generated by using randomized primers for cDNA synthesis, adapter primers were used in the first PCR step and size selection was obtained by gel electrophoresis. The same way of cDNA synthesis and normalization was also applied in other studies (Parchman et al., 2010; Wall et al., 2009). However, none of these checked for duplicated reads. Library construction and normalization protocols minimizing PCR steps and preventing the occurrence of duplicate reads would be preferable (Kozarewa et al., 2009). Nevertheless, data should always be checked for duplicated reads in order to remove them.

2. Assembly and SNPs detection:

CAP3 assembly

CAP3 uses an overlap-layout consensus algorithm for cluster construction and as such is suitable for SNP mining (Tang et al., 2008), although it is a relatively slow assembler. EST data were clustered by CAP3 with a stringency level of 95% similarity per 100 bp. The CAP3 alignment resulted in 576,882 reads that were assembled in 72,540 contigs (38.4 Mb) with an average of 8 reads per contig (Table 1). Around 26% (202,230) of the reads were singletons. The average length of the contigs was 530 bp, 274 contigs (0.38%) were less than 200 bp in length. Around 2.5% (1780) of the contigs were longer than 1 Kb, 4 contigs were longer than 2 Kb of which the longest contig was 2,800 bp (Fig. 1). A total of 10,461 reliable SNP markers were identified by QualitySNP (Tang et al., 2006).

Table 1: Comparison between CAP3 and CLC assembly results

	CAP3	**CLC**
No. Contigs	72,540	55,433
Average contig length	530 bp	555 bp
Assembled	576, 882	646,424
No. Singletons	202,230	132,688
Average reads/contig	8	11.66
No. SNP markers	10,461	2,421

CLC assembly

CLC uses the de Bruijn algorithm which is used in several assembler software packages such as Velvet (Zerbino and Birney, 2008), Oases [http://www.ebi.ac.uk/~zerbino/oases/](Oases), ABySS [http://www.bcgsc.ca/platform/bioinfo/software/abyss](ABySS), and SOAPdenovo [http://soap.genomics.org.cn/soapdenovo.html]. Similar to CAP3, a cutoff threshold of 95% was used which resulted in the assembly of 646,424 reads in 55,433 contigs (30.8 Mb) with an average of 12 reads per contig (Table 1). Around 17% (132,688) of the reads were left out as singletons. The average length of the contigs was 555 bp, 177 contigs (0.32%) were less than 200 bp in length. Around 8.5% (4709) of the contigs were longer than 1Kb, 485 contigs exceeded 2 Kb of which the longest contig was 9,420 bp (Fig. 1). A total of 2,421 SNP markers were identified by QualitySNP as reliable markers.

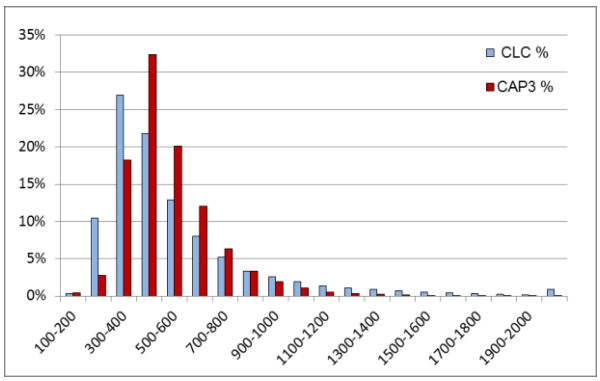


Figure 1: The distribution percentage of CLC and CAP3 contigs length (bps). The distribution percentage of contig lengths assembled by CLC and CAP3 assemblers.

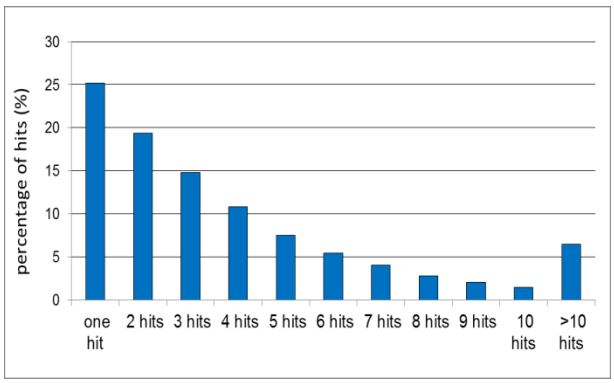


Figure2: CAP3-contigs blast vs. CLC contigs. This graph present the number of hits resulted by blasting the CAP3 contigs vs. CLC contigs. Around 25 % of the CAP3 contigs had one hit and all the rest (75%) have more than one blasting hit with CLC contigs.

A reasonable percentage of *Lilium* transcriptome was covered as could be estimated from the transcriptome size of the monocot model species rice, which has 41,000 genes with average gene length of 2,000 bp (Sterck et al., 2007). Assuming that lily has a comparable transcriptome size, the CAP3 contigs cover around 47% of the *Lilium* transcriptome while the CLC contigs cover 38%, regardless of the singletons that could be added to the total coverage.

Notable differences between the assemblers' performance were recorded. Similarly, differences in assemblers' performance were also found in another study (Feldmeyer et al., 2011). The performance of different assemblers (Velvet, Oases, and SeqMan NGen) were compared on a non-model species (snail) and showed that the assembly is strongly depend upon the assembler (Feldmeyer et al., 2011). In this study, CLC assembled more reads compared to CAP3 and also generated longer contigs with a higher average read coverage. However, CAP3 contigs generated more SNP markers and appeared to have a higher coverage in total sequence length. Both assemblers in addition to several other aligners were compared considering the number and mean length of the contigs, the assembled reads, and the assembly redundancy (Kumar and Blaxter, 2010). In contrast to our results, CAP3 and CLC performed comparable in their study. To our knowledge, there are no studies published in which the assembler's performance has been evaluated with respect to SNP retrieval. SNP markers will segregate nicely in mapping studies if the SNP is true (reliable) and if the marker is unique throughout the genome (high quality). The first step to generate reliable and high quality SNP markers is building contig in which alleles are joined and paralogs are preferably separated.
To choose the best assembler with respect to the identification of high quality reliable SNP markers for genetic mapping, we performed several tests to compare the performance of the assemblers.

3. Comparison between the CAP3 and CLC assemblies:

Assembly redundancy:

Redundancy is a main parameter in which the quality of assemblies can be evaluated. Redundancy occurs when different contigs are likely to originate from the same locus as defined by the degree of similarity (Papanicolaou et al., 2009). This is related to high numbers of differing bases which may be due to alternative splicing, multiple SNPs, InDels, and mismatches and misalignments due to homo-polymer tracts all of which show high frequencies in an outbreeding and highly heterozygote species such as in our case. The best assembler will assemble the largest number of unique sequences regardless of the number of contigs. A high redundancy of contigs is an indicator of poor assembly (Kumar and Blaxter, 2010). A pair wise comparison was performed by blasting the contigs of CAP3 *vs.* CLC-contigs with a threshold of E-20. This comparison will help verify if the differences in contigs size between the two assemblies were related to novel sequences or to the presence of repetitive and redundant contigs (Kumar and Blaxter, 2010). Results showed that only 25% of the CLC-contigs have a unique blast hit to a single CAP3 contig (Fig. 2). Similarity values of all hits exceeded E-45 except for a few cases where it ranged between E-20 and E-35 with identities above 90%.

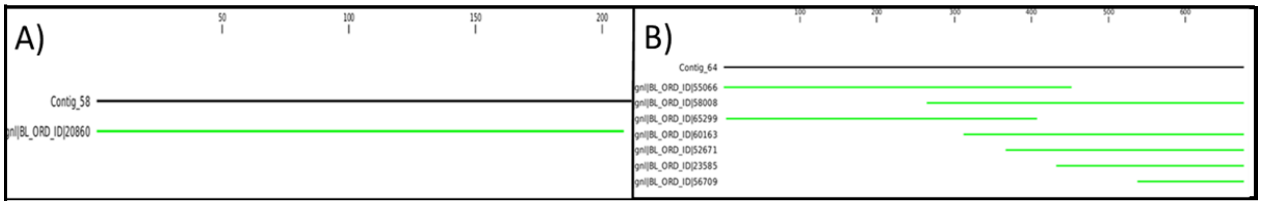


Figure3: Configuration of two examples of basting CAP3 contigs vs. CLC contigs. **A**) In this example one contig of CAP3 assembled with one of contig CLC, **B**) seven contigs of CAP3 grouped in one contig of CLC.

Visual inspection of the blast hits showed that in some cases both assemblers have mapped the same reads and constructed identical contigs (see Fig. 3A). However, in most cases several CAP3-contigs mapped to one CLC-contig (Fig. 3b) which indicates a high level of redundancy in CAP3 contigs. Blast results from blasting all CAP3 contigs among themselves confirm this (results not shown). These results indicated a significant difference in assembly performance between the two assemblers. This might be explained by the fact that the two assemblers use different approaches. The CAP3 uses OLC (Overlap-Layout-Consensus) in assembling the data (which is also used in MIRA, Newbler, and SeqMan), while CLC uses De Bruijn graph path finding (which also are used by Oases, Velvet, and ABySS). The OLC compares the overlap of whole reads at once while De Bruijn compares small stretches of base pairs (*k*-mers, 21 in our case) and combines all similar reads in one contig. This difference in algorithms may have made CLC more able to assemble reads with a high level of heterozygosity compared to CAP3 which showed to be highly discriminating. These differences between the performances of the two assemblers were not recorded in a previous study (Kumar and Blaxter, 2010). A possible explanation could be in differences in the level of heterozygosity present in the cDNA sequences between the studies. Kumar and Blaxter (2010) used the cDNA of a model filarial nematode that

has low levels of heterozygosity (Mark Blaxter, pers. comm.), while lily is an outcrossing species with a very heterogeneous breeding pool that has high level of polymorphisms (SNPs, InDels) and combined with mistakes introduced by 454 pyro-sequencing (especially in homopolymer tracts) makes the sequence reads highly heterogeneous. The level of heterozygosity within lily cultivars is around one SNP per 50 bp (calculated based on a random set the cDNA sequences), and among the four cultivars around one SNP per 26 bp. The more polymorphisms present in sequence reads, the more divergence can be detected in the performance of assemblers the effect of genome complexity on mapping was highlighted by Palmieri and Schlotterer (2009). This study showed that complex genomes, containing many gene families and paralogs are more difficult to be mapped to a reference genome compared to a compact genome. It was also recorded that SeqMan (OLC strategy) was not able to map reads (100 bp) that contain 9% computer generated variation due to the high divergence of the reads whereas this was feasible with CLC (Palmieri and Schlötterer, 2009). The differences in abilities to deal with genome complexity and heterogeneity among sequences of the assemblers indicate the importance of assembler choice.

Contigs blast to public sequence data base ESTs:

In majority of studies using NGS technologies, available data of the sequence database was used to support the assembly step since these sequences are relatively long compared to NSG sequences and more reliable since they resulted from Sanger sequencing which is still considered the gold standard in terms of sequence reliability. For lily, the number of available EST data in the sequence database is limited with 3,329 ESTs, clustering (using the default parameters of CLC) into 381 Unigenes. These Unigenes were used to compare the performance of the two assemblers by aligning the contig consensus sequences of each assembler with the 381 Unigenes and analyzing the results. The CAP3-contigs showed a total of 251 hits, 86% of them were redundant (more than one BLAST-hit) compared to CLC that showed 260 hits of which 64% showed redundancy (Fig. 4). Although these results seem comparable, there also seem to be differences here since for CLC the contigs mainly assembled adjacent to each other (Fig. 5) rather than to the same sequence stretch as was often found for CAP3-contigs. This means that several short but unique contigs of the CLC and CAP3 assembly are positioned within the EST sequence. Thus, the use of EST data from the databases to assess the performance of the two assemblers is not informative if not the two former cases (overlap or adjacent alignment) can be well defined and distinguished.

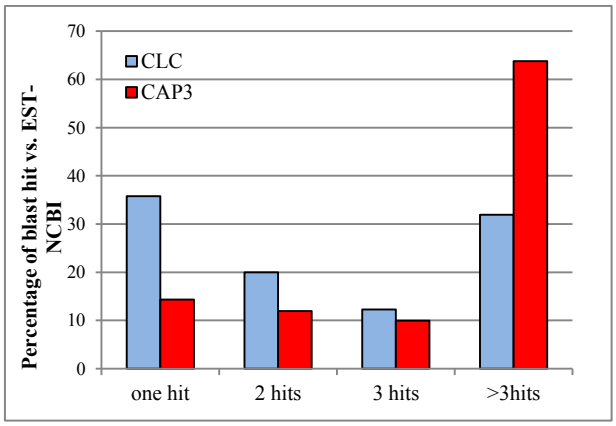


Figure 4: Blast contigs against EST-NCBI. The contigs assembled by CLC and CAP3 were blasted separately *vs.* EST-NCBI (BlastN, threshold 1E-20).

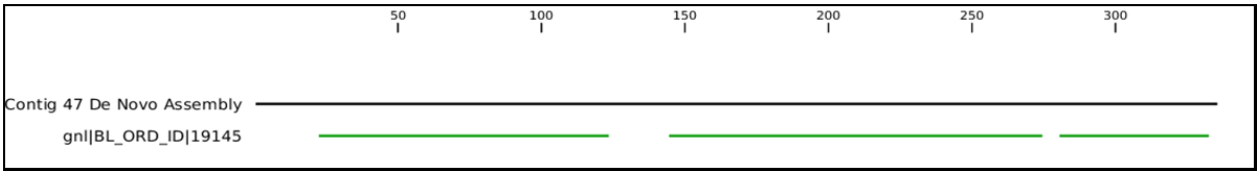


Figure 5: Configuration of blasting CLC-contigs vs. EST-NCBI. A partial matching of CLC contigs vs. EST-NCBI sequence.

Blasting generated SNPs *vs*. the contigs:

Blast results from QualitySNP selected SNPs (with 50 bp flanking sequence on each side) *v*s. the contigs from the assembly they originated, provided an additional criterion for SNP markers selection. Many species have undergone genome duplications during their evolution. Assuming paralogs are assembled in different contigs it is still possible that SNP markers selected from one of these contigs will also be present in a paralogous gene assembled in another contig. Thus, is vital to check that SNP markers only map back to the contig from which they were selected. This paralog detection is important in any study aiming to generate SNP markers for which other genetic resources are missing. Selected SNP marker sequences (101bp) of each assembler were blasted against all contigs using a threshold of E-20.

The CAP3-SNP blasting showed that only 22% of the generated 'SNP markers' (defined here as the SNP and 50 bp flanking sequence on each side) uniquely mapped to the contigs from which they were derived, 78% of the SNP markers had more than one blast hit, and 198 SNP markers had more than 10 blast hits (Fig. 6). This indicated that CAP3-SNP markers were not unique due to either a high percentage of paralogs in the *Lilium* genome or due to poor assembly and thus cannot be used for mapping studies. However, around 83% of the SNP markers generated in a CAP3 assembly of 454 transcriptome pyro-sequencing in the inbreeding species *Solanum lycopersicum*, in which the level of polymorphism is low, showed to be unique (Dr. AW van Heusden, Wageningen UR Plant Breeding, pers. comm.). From this, we can conclude that the performance of CAP3 is negatively correlated to the polymorphism level present in the genome

studied. In the case of high heterozygosity as in the present study, CAP3 software might separate alleles of highly polymorphic loci into different contigs which means that these contigs are not unique and thus it is highly risky to use them to generate SNP markers. Redundant contigs can either be related to paralogs or they can be alleles of the same locus (among the four genotypes) that were split up into different contigs. In both cases, SNP markers should not be used for mapping purposes. In case of paralogs, SNP markers will cause problems in SNP detection. In the latter case, there is a chance that SNP markers will either give no call or will work poorly because the risk of secondary SNPs close to the SNP of interest is overlooked. So, the most trustworthy SNP markers will be the ones that were generated when all alleles of the same locus are grouped in one contig. The CAP3 generated 5775 contigs (8%) that include sequences of the four cultivars, compared with 9234 contigs (17%) for CLC.

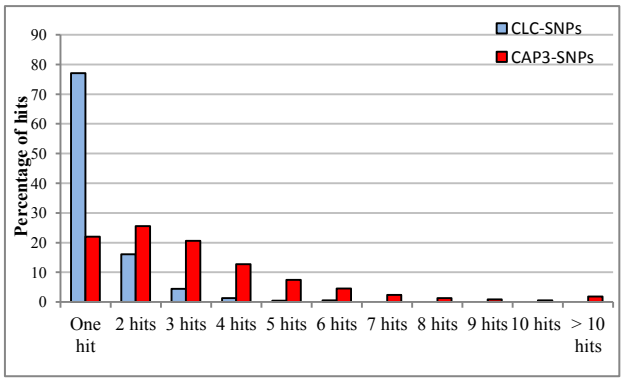


Figure 6: The percentage of hits resulted from blasting SNPs *vs*. the contigs. The CAP3-SNP (101 bp) were blasted *vs.* CAP3-contigs, and CLC-SNP (101 bp) were blasted *vs.* CLC-contigs, and the number of hits were recorded**.**

In CLC, 77% of SNP markers were unique. Only 13 SNP marker sequences had more than 5 blast hits (Fig. 6). The 22% of redundancy among CLC-SNPs can be related to the presence of paralogs assembled in different contigs. In general, a number of genes in any genome are expected to be duplicated especially in case of a huge genome like that of *Lilium*. The percentage of paralogous genes differs between species. For example, in rice around 15% to 62% were expected to be duplicated genes (Lin et al., 2008). Using a strict method of defining paralogs, the 22% of redundancy among CLC-contigs is more in line with expectations than the 78% among CAP3 contigs, especially when taking in consideration that not all paralogous genes will be expressed at the time of sampling. To check whether CLC combined paralogs in contigs, haplotype numbers were assessed. Only, 0.7% (364) of the CLC-contigs combined paralogs and contained more than the maximum expected 8 alleles (expected of 4 heterozygote diploid cultivars). The actual number of CLC-contigs with paralogs may be slightly higher but is not likely to cause high numbers of erroneous SNP markers in mapping. Thus, CLC appeared to perform reasonably well for SNP markers retrieval even with the sequence data of this highly polymorphic species. This is in correspondence with Palmieri and Schlötterer (2009) where CLC was among the two best programs for *de novo* sequence assembly. In contrast, CAP3 could not handle such high levels of heterogeneity (Palmieri and Schlötterer, 2009).

Examples of assembly differences between the two assemblers:

This step was done to visualize on the individual contig level the influence of polymorphism: on the assembly, on the SNP selection processes in each assembly, and on the Quality value D. The Quality D-value is the standard deviation of the normalized number of potential SNPs among haplotypes and it is calculated and used to assess the probability that a cluster contains paralogs (Tang et al., 2006). Randomly we selected two contigs with SNP markers. The first is a contig of CAP3 that showed a SNP marker (Contig 193) with a high Quality value (D value=0) that indicates a high quality/reliability for this SNP. High D-values indicate a high probability that a cluster contains paralogs as well as allelic sequences (Tang et al., 2008). The contig consensus of CAP3 contig 193 was blasted against the CLC-contigs and the matching contig 23548 (D=0.59, haplotypes =4) was examined. Contig 23548 has no indication of a possible SNP marker. Based on the consensus sequence, primers were developed and used to amplify the putative SNP region in the CAP3 contig 193 in the four genotypes and the obtained fragments were used in Sanger sequencing to re-sequence the putative SNP region. Sanger sequences were assembled by SeqMan (Lasergene, version 8) and then compared to contig 193 from CAP3 and contig 23548 from CLC. CAP3 contig 193 contained six reads of 'Star Gazer' with an intra SNP marker (Fig. 7a). In CLC, the same six reads together with three reads of 'Trumpet', five reads of 'White Fox' and four more reads of 'Star Gazer' formed contig 23548 (Fig. 7b). Sanger sequences (Fig. 7c) of the four genotypes confirmed that there is no 454 sequence's mistake in this locus.

The A/G SNP found within 'Star Gazer' is a reliable SNP which was also shown by CLC and confirmed by Sanger (see Fig. 7). However, the SNP was not selected as a candidate SNP marker in the CLC assembly by QualitySNP since very close SNPs (within 50 bp) were detected in the other genotypes and consequently this would not produce a general applicable SNP maker for Illumina Golden Gate (Fig. 7b, c). This example showed clearly that the CLC assembler combined reads which were separated into two contigs by CAP3 (Contig 193 and Contig 338). This indicated that CAP3 (OLC algorithm) treats alleles and homologous sequences of the same locus as paralogs belonging to other contigs or leaving them as singletons, above certain levels of polymorphism. This might explain the difference in contig and singleton number between the two assemblers. It also explains why although each contig of CAP3 contains lower levels of polymorphisms in total the assembly results in the identification of more candidate SNP markers compared with CLC that contain more reads per contig. This counter-intuitive situation is due to the higher numbers of flanking SNPs that are found in CLC-contigs and that limits the number of candidate SNPs that can be used as SNP marker in genotyping.

In the second example, contig 63 of CAP3 containing four reads; two of 'Connecticut King', one 'Trumpet' and one 'White Fox' indicated a reliable inter (G/A) SNP marker with D=0.2 (Fig. 8a). The same reads grouped together with another nine reads (one of 'Star Gazer' and eight more of 'Trumpet') in the CLC assembly (Contig 12221, D=0.6, haplotype=5). No SNP marker was selected out of contig 12221 due to presence of other close by SNPs. Surprisingly, we could not verify the SNP in the Sanger sequences generated for all 4 genotypes using primers designed on

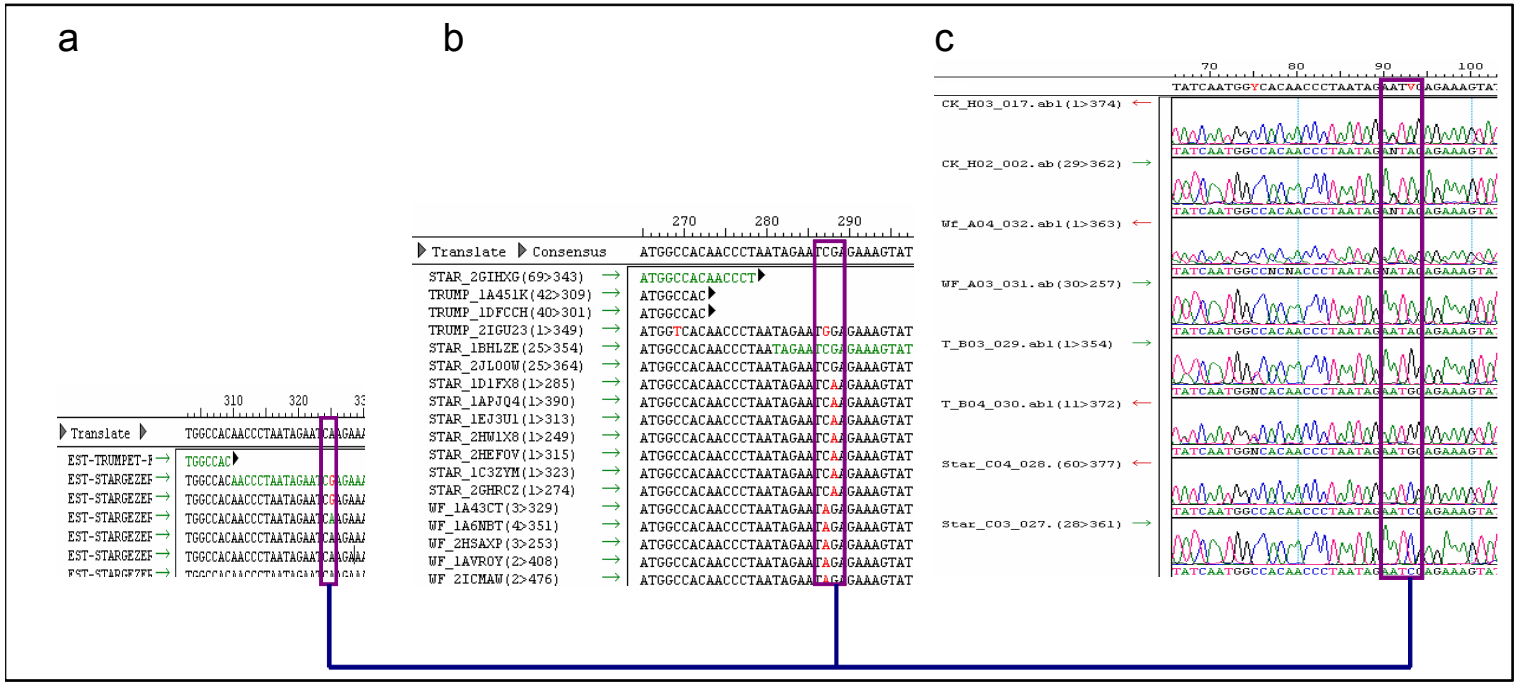


Figure 7: Comparing the same 454 sequences assembled by CAP3 and CLC from one side and with Sanger sequence of the same contig assembled by SeqMan on the other side. **a**) contig 193 resulted of CAP3 assembly, **b**) contig 23548 resulted of CLC assembly that contains all the sequences included in contig 193 of CAP3, **c**) Sanger sequencing for this contig of the four cultivars ('Connecticut King', 'White Fox', 'Trumpet' and 'Star Gazer'). Boxes indicate positions with SNPs. Lines connect the same SNP in the different contigs.

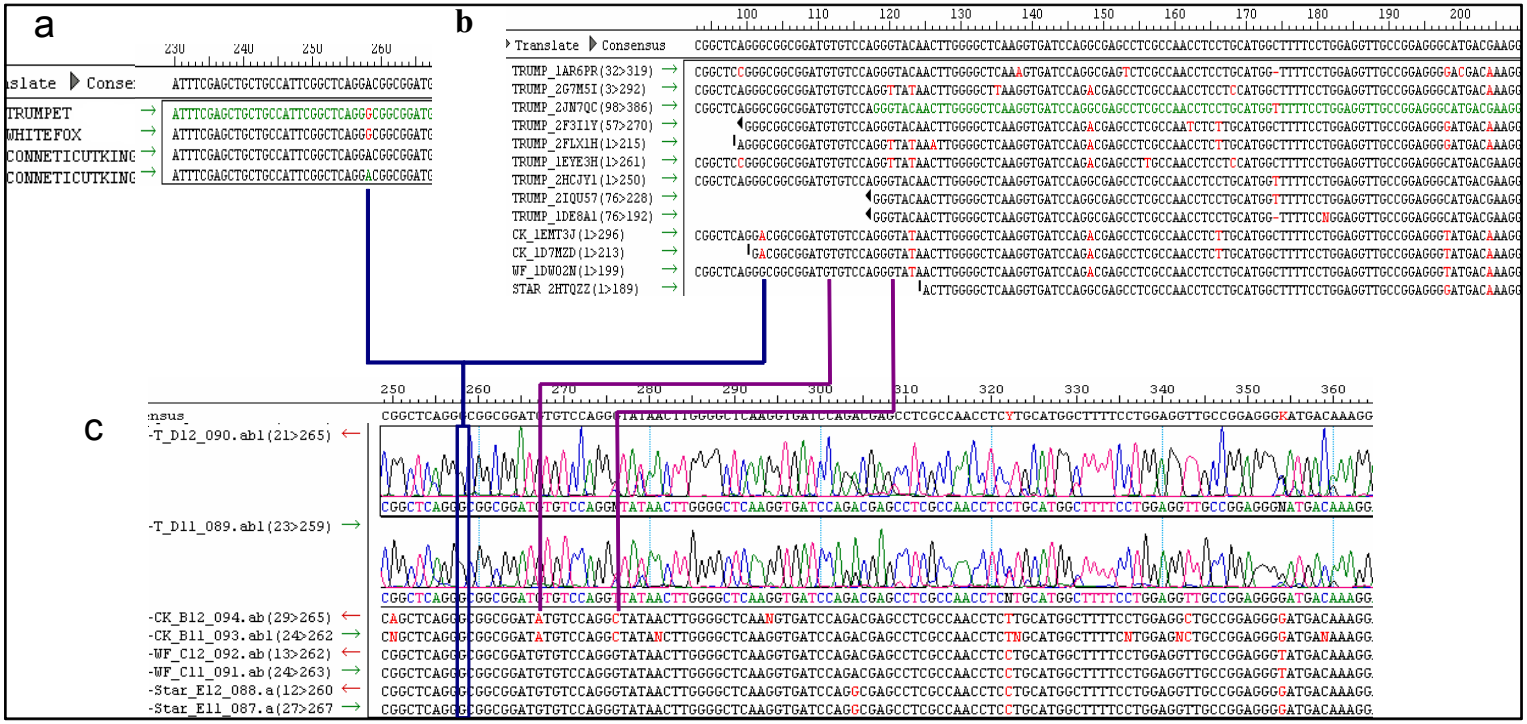


Figure 8: Comparing the same 454 sequences assembled by CAP3 and CLC from one side and with Sanger sequence of the same contig assembled by SeqMan on the other side. **a**) contig 63 resulted of CAP3 assembly, **b**) contig 12221 resulted of CLC assembly that includes all the sequences of contig 63 resulted of CAP3, **c**) Sanger sequences generated for this contig of the four cultivars ('Connecticut King', 'White Fox', 'Trumpet' and 'Star Gazer'). Lines connect the same SNP in the different contigs.

this contig (Fig. 8c). Sanger sequences indicated nucleotide (G) in all four cultivars. Another unexpected result is that the 'Connecticut King' sequences of Sanger did not match in several places with the 454 sequences (Fig. 8b and c). Moreover, CLC clearly grouped paralogs in the contig since 'Trumpet' reads show more than two alleles (Fig. 8b). By blasting the CLC contig 12221 *vs.* CAP3 contigs two hits (Contig 63 and Contig 66474) were found providing evidence that this SNP of CAP3 could not be used since it did not show a unique match to the contig from which it was selected.

This example showed that since CLC assembly adds more reads to a contig compared to CAP3, the risk of grouping paralogs into one chimeric contig might be higher. This also indicates the importance of filtering for haplotype number per cultivar and low D-values before selecting the SNP. A maximum of eight haplotypes can be expected in the assembly of four diploid heterozygous cultivars. Out of the 2,421 SNP markers selected by CLC, 4 contigs have haplotype numbers higher than 8. Number of haplotypes in contigs gives a clue on the frequency with which paralogs are incorporated into single contigs and QualitySNP uses it in SNP identification. In an ideal situation having the full genome sequence, all SNP flanking regions can be blasted to the genome and thus SNP markers can be identified for which paralogs are present. Unfortunately, for many species in which researchers would like to use the advantages of NGS technologies to develop SNP markers, whole genome sequences will not be readily available and blasting of SNP 101 bp flanking regions *vs.* contigs is the best alternative. To sum up, the main steps which are needed to generate SNP markers of a non-model species are summarized in Figure 9.

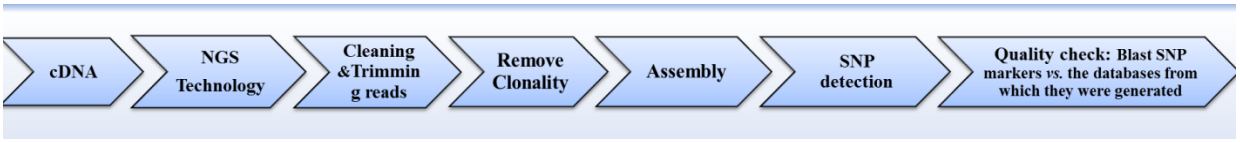


Figure 9: A scheme showing the main steps proposed to generate reliable SNP markers.

Conclusion:

SNP markers are becoming the markers of choice in genetic studies and as such for many species researchers are likely to start up SNP retrieval from NGS data. Our results clearly showed that sequence assembly and consequently the SNP markers retrieval are affected significantly by the assembler. We tested two widely used assemblers that use different algorithms. Procedures followed can be used in any species that has little genetic resources to view assembly quality. Importantly, blasting the selected SNP markers *vs.* the contigs from where they generated from (in case of missing the support information form the databases) or against the whole genome, if available, is very essential to avoid false positive SNPs. Results obtained with *Lilium* cDNAs are likely also valid in other highly heterogeneous species. There seems to be a strong correlation between the level of heterozygosity in the studied species and the performance of the assemblers.

Overall, we believe that for inbreeding species both assemblers can be used, while in an outbreeding species highly heterozygote species CLC is preferred.

Acknowledgements

Hans de Jong and Harm Nijveen are gratefully acknowledged for their constructive comments. We are thankful to Nasim Mansoori for her beneficial English editing.

Chapter 4

Generation and Analysis of Expressed Sequence Tags in the Extreme Large Genomes *Lilium* L. and *Tulipa* L.

Arwa Shahin[1,2], Martijn van Kaauwen[1], Danny Esselink[1], Jaap M. van Tuyl[1], Richard G.F. Visser[1], Paul Arens[1]

[1]Wageningen UR Plant Breeding, Wageningen University and Research Centre, P.O. Box 386, 6700 AJ, Wageningen, the Netherlands

[2]Graduate School Experimental Plant Sciences, Wageningen University

Submitted

Abstract

Bulbous flowers such as lily and tulip (*Liliaceae* family) are stunning monocot perennial herbs that are worldwide economically very important ornamental plants. A large collection of germplasm with huge genetic diversity that could serve as a basic source of different interesting traits for breeding purposes is available from both species. To speed up breeding in these two species, molecular markers for molecular assisted breeding are needed. Next generation sequencing technologies are able to generate huge amounts of sequencing data in a short time which can be implemented in all types of genetic and genomic studies. We sequenced and assembled transcriptomes of four lily genotypes ('Connecticut King', 'White Fox', 'Star Gazer', and 'Trumpet') and of five tulip genotypes ('Cantata', 'Princeps', 'Kees Nelis', 'Ile de France', and 'Bellona') using 454 pyro-sequencing technology. Successfully, we developed a first set of 81,791 unigenes with an average length of 514 bp for tulip, and enriched the very limited number of 3,329 available ESTs (Expressed Sequence Tags) for lily with 52,172 unigenes with an average length of 555 bp. These unigenes were used to identify SNPs (single nucleotide polymorphism) markers suited for high throughput genotyping purposes. A total of 2,421 and 1,079 SNP markers were generated for lily and tulip respectively using very strict conditions (no SNPs in the flanking 50 bps on either side of the SNP were allowed). These numbers increased to 5,614 and 3,525 SNP markers for lily and tulip respectively if one SNP was allowed in the flanking regions. Orthologous genes between lily and tulip were identified (10,913 unigenes), and together with the whole set of unigenes generated for lily and tulip they were annotated and described according to Gene Ontology (GO) terminology. We also screened the sequences of both plants for SSR markers and identified 882 and 1,379 SSR markers for lily and tulip respectively. In the orthologous sequences between the two species, 229 common SNP and 140 common EST-SSR markers were identified, which can be applied in comparative genomic studies between the two species later.

Introduction

Lily and tulip (*Liliaceae* family) are monocot perennial herbs that have unsurpassed beauty and great commercial value. They are also very interesting from an evolutionary point of view since both species have very huge genomes, which make genetic and genomic studies challenging. The two species are comparable in several aspects: both are monocots, bulbous flowers, with a very large genome size (1C=25 GB for tulip, and 36 GB for lily), have 12 large chromosomes, and have a long growth cycle (2-3 years for lily and 5-6 years for tulip). And importantly, for both species genetic resources are limited.

Genus *Lilium*, includes around 100 species and more than 9,400 cultivars (International Lily register, http://www.lilyregister.com/) which are taxonomically classified into seven sections: *Martagon, Pseudolirium, Lilium, Archelirion, Sinomartagon, Leucolirion*, and *Oxypetala* (Comber, 1949; De Jong, 1974). The wild species within each section are relatively easy to cross

and the hybrids are fertile (McRae, 1990; Van Tuyl et al., 2002). The inter-specific hybrids within sections *Leucolirion, Archelirion*, and *Sinomartagon* represent the most important groups for breeding: Longiflorum (L), Trumpet (T), Asiatic (A), and Oriental (O) hybrid groups. An extensive number of cytogenetic studies explored the karyotypes of lily (Bach Holm, 1976; Lim et al., 2001; Marasek et al., 2004). Meiosis of inter-specific hybrids and cytological maps of three complete genomes of lilies (L, A, O) based on the recombination sites in the BC progeny of two interspecific hybrids (Khan et al., 2009) were studied. On the other hand, genetic mapping of lily has not yet been well studied. The currently available genetic maps constructed using dominant markers (AFLP 'Amplified Fragment Length Polymorphism', NBS 'Nucleotide Binding Site', and DArT 'Diversity Arrays Technology') are not well saturated (Shahin et al., 2011). The available EST data in the sequence database is very limited with only 3,329 ESTs deposited.

Genus *Tulipa* L. (2n=2x=24), comprises about 100 species (Bryan, 2002) that are taxonomically classified into two subgenera: *Tulipa* and *Eriostemones* (Van Raamsdonk and De Vries, 1992; Van Raamsdonk and De Vries, 1995). Subgenus *Tulipa* is subdivided into five sections named: *Tulipa*, *Eichleres*, *Tulipanun*, *Kolpakowskianae*, and *Clusianae*. The commercial cut flower assortment of tulips consists mainly of cultivars from *Tulipa gesneriana* (section *Tulipa* and *T. fosteriana* (section *Eichleres*) (Van Creij et al., 1997). Tulip chromosomes are identified and karyotyped (Mizuochi et al., 2007). So far, there are no genetic maps or molecular markers published for tulip, and additionally no ESTs are found in the databases for this species.

Breeding in these two species is limited by their long juvenile phase, and it becomes more challenging when breeding for quantitive traits. Both species suffer from *Fusarium*, *Botrytis*, and tulip breaking virus diseases (Lim and Van Tuyl, 2006; Van Tuyl and Van Creij, 2006). *Fusarium* resistance in lily is known to be controlled by six putative QTLs (Shahin et al., 2011). Developing user friendly, efficient, transferable, and co-dominant markers such as single nucleotide polymorphism (SNPs), and simple sequence repeat (SSRs) markers that can be implemented in molecular assisted breeding (MAB) applications will help to speed up breeding in these two species (Dubey et al., 2011). Transcriptomes or ESTs analysis is one of the most efficient and effective approaches for identification of candidate genes and assist in new molecular marker development such as SNPs (Narina et al., 2011) and SSRs (Morgante and Olivieri, 1993). Additionally, generation of ESTs would be extremely desirable for transcriptome characterization in lily and tulip, from which a general description of genes present and expressed can be identified. As breeding is time consuming for both species (both have a long juvenile phase), transfer of developed genetic knowledge between the two species will help to speed up and support the breeding process. To achieve that, the study of micro-synteny between lily and tulip (orthologous genes), and common markers that can be mapped in both species will be of great help.

Recent studies have shown that next generation sequencing technology can be an effective tool to generate huge amounts of sequence data in a short time which can be implemented in all types of

genetic and genomic studies such as: transcriptome characterization, molecular marker development (Blanca et al., 2011; Huang et al., 2011; Narina et al., 2011), ecological genetics (Wheat, 2010), and evolutionary studies (Gilad et al., 2009). With the purpose of generating the first broad survey of genes in lily and tulip, we sequenced and assembled transcriptomes of four lily and five tulip genotypes using 454 pyro-sequencing. The sequence assemblies were used to identify a set of SNPs suited for high throughput genotyping purposes, and to screen for SSR markers. Also, orthologous genes between lily and tulip were identified, and the whole set of ESTs generated for lily and tulip, were annotated and described according to GO terminology. Common markers that can be genotyped and mapped in both species were identified.

Materials and Methods

Plant material

Four lily genotypes that represent the four main hybrid groups of genus *Lilium* were used for sequencing: cv. 'Star Gazer' (Oriental), breeding line 'Trumpet 061099' (Trumpet), cv. 'White Fox' (Longiflorum) and cv. 'Connecticut King' (Asiatic). Five tulip cultivars were used for sequencing: cv. 'Cantata' and cv. 'Princeps' belonging to *T. fosteriana* (*Eichleres* section) and cv. 'Bellona', cv. 'Kees Nelis', and cv. 'Ile de France' belonging to *T. gesneriana* (*Tulipa* section). Young leaves (500mg) were collected and kept at -80°C upon RNA isolation.

Methodology

RNA isolation, cDNA library preparation, 454 sequencing procedures, and the assembly using CLC assembler were described in Chapter 3 for lily and the same steps were applied for tulip cultivars.

Briefly: RNA was isolated using the Trizol protocol (Invitrogen. Carlsbad, CA, USA) and purified using the RNeasy MinElute kit (Qiagen, Hilden, Germany). The cDNA synthesis, normalization of the cDNA, and adaptor ligation for GS FLX Titanium sequencing were performed by Vertis Biotechnologie AG (Freising, Germany). In short, 45 ug of total RNA of each of the samples was treated with DNase and then primed with a 6 nucleotide randomized primer for first strand cDNA synthesis. Next, 454 adapters A and B with a 6 nucleotides barcode for each cultivar were ligated to the 5' and 3' ends of the cDNAs. The cDNAs were subjected to two steps of PCR using a proof reading enzyme before and after normalization. Normalization was carried out by one cycle of denaturation and re-association of the cDNAs, and column purification. For 454 sequencing, cDNAs in size range of 500 – 600 bp were eluted from preparative agarose gels.

cDNA sequence processing and assembly

The cDNA libraries, with the A and B 454 adapters at both ends, were mixed in equal concentrations and sequenced on a Life Sciences GS-FLX Titanium according to standard procedures (454 Life Sciences) at Greenomics (Wageningen, the Netherlands).

Using CLC genomics workbench software (CLC bio, Denmark, http://www.clcbio.com/), the 3' and 5' adapter sequences were trimmed. Low quality bases (1 base at the 3' end and 15 bases from the 5'end, other low quality terminal bases with a 0.05 threshold) were also removed, and the maximum ambiguous nucleotides present in the fragment were set to 2. Only fragments between 100-800 bp were kept for further analysis. Longer and shorter sequences were discarded. CD-HIT (Li and Godzik, 2006) was used to remove PCR duplicates (clonality) with a threshold of 98% similarity. The *de novo* assembly using CLC was done using the following parameters: conflict resolution (vote), similarity 95%, and alignment mode (global, do not allow InDels). Also, the total number of SNPs was calculated using the default parameters of CLC assembler.

The contigs (non-redundant sequence or unigenes) were constructed: for each genotype separately, for the four lily genotypes together, for the five tulip genotypes together, for *T. fosteriana* cultivars ('Cantata' and 'Princeps') together, and for *T. gesneriana* cultivars ('Bellona', 'Kees Nelis' and 'Ile de France') together.

SNP marker detection

All contigs/unigenes resulting from CLC were submitted to an updated version of QualitySNP (Tang et al., 2006; http://www.bioinformatics.nl/tools/snpweb/) to detect single nucleotide variants (SNPs). The SNPs were chosen based on the following criteria: high quality sequence, not within or adjacent to a homopolymeric tract, at least 2 reads of each allele and 50 bp of flanking sequence on each side. The resulting SNP regions (50 bp flanking SNP on each side) were compared against all contigs using BlastN with Expectation value E lower than -20. Only SNPs, which mapped uniquely to the contig from which they were selected, were retained.

SNP marker selection was done twice using different stringency conditions. The first time with a flanking region of 50 bps on either side free of SNPs, and the second time we allowed one SNP in flanking region. This was done since different high throughput genotyping technologies require different conditions. To ensure high quality of SNP markers, D value was limited to (0-0.5) which reduces the probability that an assembled cluster contains paralogs.

Mining for microsatellites

Microsatellites were searched using MISA (Thiel et al., 2003) which identifies perfect compounds and interrupted microsatellites. The criteria for selection of microsatellites were a minimum of six repeats for di-nucleotide motifs and five repeats for tri-, tetra-, penta-, and hexa-nucleotide motifs were used. Primer 3.0 software (Rozen and Skaletsky 2000) was used to design primers flanking the putative SSRs. The input criteria for Primer 3.0 were: a primer length of 17–30 bp, a GC content of 20–80%, and an estimated amplicon size of 50–300 bp.

Orthologous sequence, gene annotation and gene ontology identification

All tulip unigenes were blasted (BlastN, 1E-20) *vs.* lily unigenes and all unigenes that showed similarity between the two species were selected for further analysis and referred to as orthologous sequences.

Next, lily and tulip's unigenes (Lily-All and Tulip-All) and the orthologous sequences were annotated by blasting (BlastX) to the databases (non-redundant protein sequences-nr) using Blast2Go V.2.4.9 software (Conesa et al., 2005) with an E-value of 1E-15. Blast2Go is an automated tool for the assignment of gene ontology terms and was designed for use with novel sequence data. Distribution of genes in each ontology category was examined and percentage of unique sequences in each of the assigned GO terms: biological process, molecular function, and cellular component were computed and presented.

Identification of common SNP and SSR markers within and between the two species

Mapping populations are available for both species. In lily, an inter-sectional F1 population (100 progenies) resulting from a cross of cv. 'White Fox' (section *Leucolirion*) with cv. 'Connecticut King' (section *Sinomartagon*) (Khan, 2009). In tulip, a F1 inter-sectional population (around 100 progenies) resulting from a cross of cv. 'Kees Nelis' (*T. gesneriana*) with cv. 'Cantata' (*T. fosteriana*). To link the two species and be able to transfer information from one species to another, common markers that can be mapped in both lily and tulip populations are needed. The common markers in this study refer to the markers generated of the orthologous sequences between and lily and tulip (might or might not present the same polymorphism, *i.e.*: each represents certain polymorphism within the same orthologous sequence), and thus their mapping position can be used to study synteny between the two species. Similarly, common markers within each species (between 'White Fox' and 'Connecticut King'; and between 'Kees Nelis' and 'Cantata') were identified.

Results and Discussion

EST sequencing and assembly

We performed 454 GS FLX Titanium pyro-sequencing on nine cDNA libraries constructed from leaves of four lily genotypes ('Connecticut King', 'White Fox', 'Star Gazer', and 'Trumpet 061099'), and five tulip genotypes ('Cantata', 'Princeps', 'Ile de France', 'Kees Nelis', and 'Bellona'). The number of sequenced reads obtained varied between 139,480 reads for 'Connecticut King' and 592,034 reads for 'Kees Nelis' (Table 1). The portions of sequence reads that were retained for assembly after filtration ranged between 67% and 75% (Table 1) which was somewhat higher than the percentage reads retained after quality filtration of 454/Sanger data of *Eucalyptus* (60.7%) (Grattapaglia et al., 2011), and close to the 79% of *Pinus contorta* (464,896 retained after filtration of 586,732 reads generated by 454 sequencing in pine, Parchman et al., 2010). Average read length ranged between 278 bp for 'Bellona' and 389 bp for

'Cantata' (Table 1). These results were comparable (and even better in some genotypes) with that obtained in other studies like Blanca et al. (2011) where the processed reads of cucurbit retained after trimming was 64% with an average read length of 321 bp. After filtration, the remaining reads were used for *de novo* assembly using CLC assembler.

Table1: Statistics of 454 sequence assembly for four lily and five tulip genotypes

Genotype	No. reads	No. reads after filtration	Avg. read length bp	No. assembled reads	Singletons	No. unigenes	Avg. EST length bp
Connecticut King	139,480	104,323(75%)	336	77,097(74%)	27,226(26%)	14,773	615
White Fox	326,539	221,597(68%)	338	182,393(82%)	39,204(18%)	21,898	663
Star Gazer	374,240	255,081(68%)	341	202,707(79.5%)	52,374(20.5%)	24,700	688
Trumpet	442,476	299,655(69%)	343	241,782(81%)	57,873(19%)	26,075	694
Lily-All	1,282,735	880,656(69%)	340	471,378(53.5%)	409,278(46.5%)	52,172	555
Cantata	310,973	207,229(67%)	389	158,007(76%)	49,222(24%)	17,646	625
Princeps	316,372	211,380(67%)	386	165,282(78%)	46,098(22%)	17,007	632
Kees Nelis	592,034	407,392(69%)	281	303,558(74.5%)	103,834(25.5%)	38,716	559
Ile de France	263,175	185,464(70%)	283	125,293(67.6%)	60,171(32%)	24,557	517
Bellona	221,334	149,768(67%)	278	109,309(34%)	40,459(27%)	14,325	522
T. fosteriana	627,345	418,609(67%)	388	293,043(70%)	125,566(30%)	24,713	629
T. gesneriana	1,076,543	742,624(69%)	281	536,776 (74%)	205,848(28%)	54,575	557
Tulip-All	1,703,888	1,378,898	314	827,772(60%)	551,126(40%)	81,791	514

Currently, a total of 3,090 lily's ESTs are available in the nucleotide sequence databases generated from *Lilium formosanum* (1324), *L. longiflorum* (991), Oriental hybrids (565), and *L. regale* (210). These ESTs were clustered into 381 unigenes (see Chapter 3). In this study, we generated 52,172 consensus sequences (non-redundant sequences or unigenes) representing gene fragments from the four main groups of *Lilium*. We, also, generated for the first time 81,791 unigenes for tulip representing the two main groups of commercial tulips: *T. fosteriana* (two cultivars) and *T. gesneriana* (three cultivars). Overall, the number of lily unigenes generated in this study is comparable to that obtained in previous transcriptome analysis such as in cucurbit (49,610 unigenes generated out of two cultivars and three different tissues, Blanca et al., 2011), and in *Eucalyptus* (48,973 unigenes generated out of mixed Sanger/454 databases of six species) (Grattapaglia et al., 2011). The number of tulip unigenes is at the high end. It is, however, important to keep in mind that number of generated unigenes does not reflect number of genes. Fragments of one gene could be assembled in different unigenes due to: short unigenes length (range of 500 to 700 bp) compared with the average gene length (2 Kb), missing overlap among unigenes which might be related to the not fully unbiased cDNA synthesis step in sequence library construction using random hexamer primers, or orthologous sequences among genotypes are assembled into different contigs due to high genetic divergence among different genotypes.

Running assembly for the four lily genotypes together (Lily-All assembly) resulted in dramatic increase in singletons number. Number of singletons was expected to be either additive of the four assemblies of four genotypes separately, or less since some singletons of different genotypes can be grouped together. In Lily-All assembly, however, number of singletons (409,278 reads) was far higher than the sum of the singletons of the separate assemblies for four lily genotypes

(176,677 reads). Similar observation was found in tulip in which 551,126 reads were left as singletons in Tulip-All assembly while the sum of singletons from the five genotype assemblies separately was around 300,00 reads. On the other hand, by assembling the two genotypes of *T. fosteriana,* singleton number (125,566) was close to the summed singletons number of each assembly separately (95,320 reads). Similarly, *T. gesneriana* assembly resulted in 205,848 singletons and the sum of singletons in the three separate assemblies was 204,464. This could be explained by similarity parameter that was defined for assembly step (95% in this study). Clustering reads of different species or sections, in addition to the presence of alternative splicing and homopolymeric track variations will increase divergence between reads. Thus, assemblers may fail to assemble all possible reads and leave them as singletons. This phenomenon was less observed in closely related genotypes where the genetic divergence is less. This reflects the importance of setting correct parameters for each assembly. However, since paralogs may be present in these sequences, using a less stringent similarity parameter cannot be implemented without careful deliberation. Given the number of unigenes and markers discovered using the current settings in this study, no further optimization of assembler parameter settings were examined.

Remarkably, number of unigenes also increased in Lily-All and Tulip-All assemblies compared with the separate assemblies of genotypes involved in this study; even though, more singletons were left out (Table 1). This either because different sets of genes being sequenced of different genotypes, or and that orthologous sequences among the genotypes tend to split up into different contigs due to increase the level of heterogeneity among the genotypes (Shahin et al., 2012).

An estimation of the transcriptome coverage of lily and tulip genotypes was calculated (Table 2). There is no information about the total size or the number of genes in lily and tulip. Therefore, the transcriptome size was assumed to be similar to other monocot species such as rice. Gene space was estimated to be around 82 Mb in rice (41,000 genes with an average gene length of 2 Kb, Sterck et al., 2007). Gene coverage of each genotype was calculated based on the total number of bases generated (assembled sequences and singletons) as a percentage of the estimated gene size (82 Mb similar to rice). In lily, gene coverage varied between 26% in ‘Connecticut King’ and 46% in ‘Trumpet’. In tulip, the lowest coverage was in ‘Bellona’ and the highest was in ‘Kees Nelis’ (23 and 63 %, respectively). *T. gesneriana* genotypes seem to cover the entire gene space although two-thirds was derived of singletons (Table 2). Gene coverage of Lily-All and Tulip-All was not calculated since singletons might not be unique as was explained previously. The large number of unigenes generated and the good coverage of the transcriptome for both species, from a single 454 run of cDNA libraries, constructed out of one tissue and in a single growing stage, shows the high efficiency of next generation sequencing technology.

Table 2: The estimated percentage of transcriptome coverage for each genotype was calculated based on the number of genes and average gene size of rice. This percentage was not calculated for Lily-All and Tulip-All due to the uncertainty of having unique singletons. The total numbers of SNPs detected within each genotype and among different genotypes were calculated as percentage compared to the total base pairs of assembled sequences.

Genotype(s)	Assembled Sequences (MB)	Singletons (MB)	Total (MB)	Transcriptome Coverage %	Total SNPs	SNP %
Connecticut King	10	11.2	21.2	25.8	10,348	0.1
White Fox	14.5	13.2	27.7	33.8	13,179	0.1
Star Gazer	17	17.9	34.9	42.6	19,725	0.12
Trumpet	18	20	38	46.3	24,282	0.14
Lily-All	29	143	~	~	131,808	0.46
Cantata	11	19.5	30.5	37.2	22,943	0.21
Princeps	10.8	18	28.8	35	24,534	0.23
Kees Nelis	21.6	30	51.6	63	28,325	0.13
Ile de France	12.7	17	29.7	36	9,910	0.08
Bellona	7.5	11	18.5	22.6	8,311	0.11
T. fosteriana	16	50.7	66.7	81.3	39,311	0.24
T. gesneriana	30.4	60	90.4	110	48,308	0.16
Tulip-All	42	182	~	~	104,338	0.48

Polymorphism in lily and tulip

An estimation of the overall sequence polymorphism rate (SNPs) is complicated by uncertainty over the actual complexity of transcriptomes (presence of splicing variants, paralogs, orthologous sequences.*etc*.) being analyzed. A general estimation of polymorphism rate in lily and tulip species was calculated (*i.e.* percentage of the total number of identified SNPs compared to the total base pairs assembled, Table 2). In lily, percentages of SNPs within transcriptome of each genotype were comparable (0.1% in 'Connecticut King' and 'White Fox', and it increased slightly in both 'Star Gazer' and 'Trumpet' to reach 0.12% and 0.14%, respectively). This adds up to one SNP every 1000 bp approximately. The percentage of SNPs among four lily genotypes of three sections (*Archelirion*, *Sinomartagon*, and *Leucolirion*) reached 0.46%, approximately 1 SNP every 200 bp, calculated in the same way. *Sinomartagon* section resulted from crossings among 10 species, which may suggest that the level of heterozygosity/ heterogeneity is expected to be high in this section compared with *Archelirion* section where the cultivar groups were derived from crossing among 3 species, and with *Leucolirion* section that were derived from one species only. This might mean that the level of heterozygosity/ heterogeneity is expected to be higher in 'Connecticut King' compared to 'Star Gazer' and 'White Fox', which was not confirmed by our data.

In tulip, a clear difference in polymorphism percentage between *T. gesneriana* genotypes (the lowest percentage was in 'Ile de France' with 0.08%, and the highest was in 'Kees Nelis' with 0.13%) and *T. fosteriana* genotypes (0.21% and 0.23% for 'Cantata' and 'Princeps', respectively) was found. An average of 1 SNP per 500 to 1000 bp was estimated in tulip. The polymorphism among the three genotypes of *T. gesneriana* was lower than the polymorphisms between the two genotypes of *T. fosteriana* (0.16% and 0.24%, respectively). In other words, *T. fosteriana* is at least one times more diverse than *T. gesneriana*. Polymorphism percentage among tulip genotypes increased to that detected among the four genotypes of lily (0.48% compared with

0.46% for lily), one SNP every 200 bp in tulip. These percentages are comparable to that detected in the out crossing species maize where the SNP frequency ranged between one SNP per 600 bp to one per 100 bp of aligned sequences depending on the number of reads per contigs (SNP frequency increased with increasing the number of reads per contig) (Batley et al., 2003).

SNP marker detection

QualitySNP software was used to identify single nucleotide polymorphisms by comparing reads within each cluster. The SNPs were declared to be reliable when at least two individual reads in the cluster have a variant allele and at least two reads have the allele of the consensus (Tang et al., 2006). We analyzed only single nucleotide polymorphisms (SNPs) and excluded all InDels due to the fact that 454 has serious problems with mono-nucleotide tracts and may introduce InDels without biological significance frequently. The number of clusters (unigenes) that contain at least one putative SNP ranges from 26% in 'Ile de France' to 47% in 'Princeps' (Table 3). The SNP markers were filtered using two sets of conditions. In the first set of conditions, any SNP in the 50 bp flanking regions of the target SNP was excluded, next, percentages of SNP markers (that follow this condition) compared to the total number of unigenes that have at least one SNP were calculated. The highest percentage in lily was in 'Connecticut King' (9.4%), while the other three cultivars showed slightly lower percentages (around 6%). In tulip, the highest percentage was for *T. fosteriana* cultivars (10%), which was two times more than that calculated for *T. gesneriana* genotypes (5%). In the second set of conditions, SNP marker selection allowed a secondary SNP in the flanking regions. Allowing a secondary SNP increased percentages of identified SNP markers in all genotypes (2 to 3 fold more, Table 3). A large number of SNP markers was generated in each genotype (ranging between 1,1171 and 2,075 SNP markers in lily and between 535 and 2,510 SNP markers in tulip, when a secondary SNP was allowed) compared with around 572 SNP markers generated in *Eucalyptus* when no control on the flanking SNPs was applied (Grattapaglia et al., 2011).

Table 3: SNP markers identification: with 50 bps flanking sequences free of secondary SNP, and with one secondary SNP allowance.

Genotype(s)	No. unigene	No. Unigenes has at least one SNP	SNP markers (no secondary SNP)	SNP markers (one secondary SNP)
Connecticut King	14,773	4,309(29%)	406(9.4%)	1,171(27%)
White Fox	21,898	9,261(42%)	558(6%)	1,292(14%)
Star Gazer	24,700	10,024(41%)	730(7%)	2,026(20%)
Trumpet	26,075	11,298(43%)	607(5%)	2,075(18%)
Lily-All	52,172	24,613(47%)	2,421(10%)	5,614(23%)
Cantata	17,646	7,456(42%)	722(10%)	2,371(32%)
Princeps	17,007	7,587(45%)	690(10%)	2,510(33%)
Kees Nelis	38,716	13,832(36%)	595(4.3%)	1,646(12%)
Ile de France	24,557	6,347(26%)	310(5%)	776(12%)
Bellona	14,325	4,476(31%)	223(5%)	535(12%)
T. fosteriana	24,713	11,787(48%)	1,002(8.5%)	3,265(28%)
T. gesneriana	54,575	20,661(38%)	822(4%)	2,033(10%)
Tulip-All	81791	31,042(38%)	1,079(3.4%)	3,525(11%)

The percentage of unigenes that have SNPs exceeded 40% for all lily genotypes except for 'Connecticut King' (Table 3). In tulip, the percentage of unigenes that have SNPs in *T. fosteriana* genotypes also exceeded 40%, while in *T. gesneriana* genotypes lower percentages were found (26% in 'Ile de France'). These results were comparable to those detected in *Eucalyptus* (40%) (Grattapaglia et al., 2011), and also they fit with the previously observed higher polymorphism rate in *T. fosteriana* compared to *T. gesneriana*.

Mining for microsatellites

We screened lily and tulip unigenes for the presence of SSRs, and analyzed their nature and frequency. Percentages of EST-SSR (compared to the total number of unigenes) found in lily genotypes were comparable with each other (around 2.7%) except for 'Connecticut King' that showed a lower percentage (1.9%) of EST-SSR in unigenes. In tulip, percentages of EST-SSR in unigenes were similar in *T. fosteriana* genotypes ('Cantata' and 'Princeps', around 4%), and similar in *T. gesneriana* genotypes ('Bellona', 'Ile de France', and 'Kees Nelis', around 2%), although lower in *T. gesneriana* compared with *T. fosteriana* genotypes. Percentages of SSRs found for lily (1,387 SSR, 2.7%) and tulip (2,029 SSR, 2.5%) were higher than results from *Medicago truncatula* in which a total of 401 out of the 184,599 ESTs contained SSRs (0.2%) using the same criteria (Cheung et al., 2006), comparable to grape and barley (3 and 2.8% respectively) (Huang et al., 2011; Varshney et al., 2006), and lower than pigeonpea (7.6 %) (Dutta et al., 2011).

Table 4: SSR motif description in lily and tulip. The total number and the percentage of SSR motif were calculated according to the total number of unigenes.

SSR motif	No. unigene	Total No. SSR	di-	Tri-	Tetra-	Penta-	Hexa-
Connecticut King	14,773	271 (1.9%)	85(31%)	161 (59%)	4	6	15
White Fox	21,898	603(2.8%)	216(36%)	301(50%)	51	12	23
Star Gazer	24,700	735(3%)	299 (41%)	330 (45%)	66	13	27
Trumpet	26,075	745(2.8%)	312 (42%)	341(46%)	50	17	25
Lily-All	52,172	1,387 (2.7%)	632 (46%)	583 (42%)	105	33	34
Cantata	17,646	696(3.9%)	168 (24%)	449 (65%)	30	9	40
Princeps	17,007	683(4%)	146(21%)	468 (69%)	28	11	30
Kees Nelis	38,716	881(2.3%)	262 (30%)	491 (56%)	58	19	51
Ile de France	24,557	521(2%)	140 (27%)	317 (61%)	33	12	19
Bellona	14,325	302(2%)	80 (28%)	184 (64%)	9	11	18
T. fosteriana	24,713	957 (3.9%)	216(23%)	642(67%)	45	15	39
T. gesneriana	54,575	1,302 (2.9%)	393(30%)	719(55%)	95	35	60
Tulip-All	81,791	2,029(2.5%)	609 (30%)	1160 (57%)	128	47	85

The di-, tri-, tetra-, and hexa-nucleotide repeats were looked up in each set of unigenes (Table 4). In both species, the most frequent repeat motif is AG/CT for di-nucleotide repeats and CCG/CGG for tri-nucleotide repeats. Similar results were found in barley (Thiel et al., 2003) which is also a large genome sized monocot. More than 86% of the identified EST-SSRs in lily and tulip are di- and tri- nucleotide repeats. In lily, the relative amounts of di- and tri-nucleotide repeats varied with equal amounts in 'Star Gazer' and 'Trumpet' (around 40 % for each repeat type) while in 'Connecticut King' and 'White Fox' tri-nucleotide repeats were more abundant than di-

nucleotide repeats (Table 4). In tulip, tri-nucleotide repeats were around two fold more abundant than di-nucleotide repeats (Table 4). This finding in tulip is in agreement with previous findings in grape and castor bean (Huang et al., 2011; Qiu et al., 2010). The dominance of tri-nucleotide repeats in transcriptome regions was expected as a result of a selection against possible frame shift mutations of one amino acid (three nucleotides) (Huang et al., 2011), while di-nucleotide repeats are dominant in the 5'- and 3'-UTRs (Dutta et al., 2011). This might suggest that the sequences of both 'Star Gazer' and 'Trumpet' that showed a comparable percentage of di-and tri-nucleotide contain large amounts of UTR regions.

The number of compound SSRs identified in lily and tulip (Table 5) was far more than that identified previously in *Lilium japonicum* (3 compound SSRs, Kawase et al., 2010). Unigenes with more than one SSR repeat, and number of SSRs for which primers could be designed are given in Table 5. SSR primer pairs develop for cultivars ranged between 64-71% for lily and 65-73% for tulip. These percentages are comparable to that observed in pigenopea (*Cajanus cajan* L., in which 2,877 PCR primer pairs were designed out of a total of 3,771 SSR identified, 76%, Dutta et al., 2011). More SSR-EST of lily and tulip are expected to be identified by running SSR mining using singleton reads.

Table 5: The SSRs identification in lily and tulip unigenes. SSR present in compound formation, and unigenes have more than one SSR motif were calculated. EST-SSR markers with primers designed and their percentages according to the total number of identified SSRs were calculated.

Genotype(s)	No. of SSRs	Nr. SSRs present in compound formation	No. unigenes has more than 1 SSR	Nr. SSR with primers designed
Connecticut King	271	13	11	193(71%)
White Fox	603	22	26	411(68%)
Star Gazer	735	44	39	478(65%)
Trumpet	745	35	44	515(69%)
Lily-All	1,387	94	89	882(64%)
Cantata	696	30	43	506(73%)
Princeps	683	30	44	487(71%)
Kees Nelis	881	35	47	590(70%)
Ile de France	521	14	20	347(67%)
Bellona	302	15	14	197(65%)
T. fosteriana	957	38	54	692(72%)
T.gesneriana	1,302	49	64	877(67%)
Tulip-All	2,029	95	118	1,379(70%)

Orthologous sequences, gene annotation and gene ontology identification

Orthologous sequences between lily and tulip were identified by BlastN tulip unigenes *vs.* lily unigenes with a cut-off 1E-20. Blasting resulted in 10,913 orthologous sequences (Fig.1). These orthologous sequences, together with lily and tulip unigenes, were annotated.

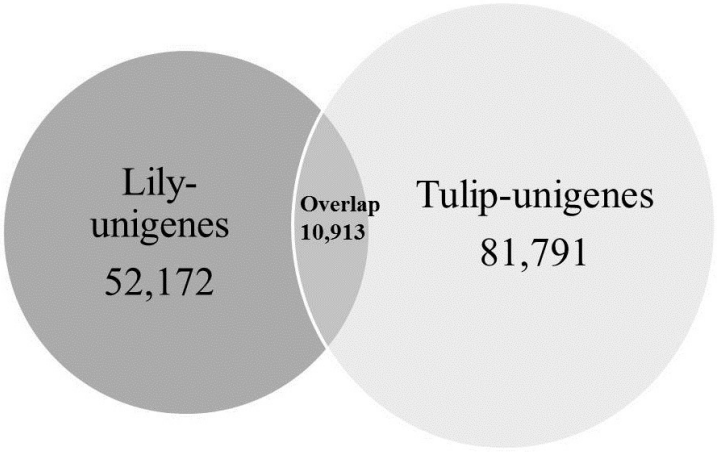


Figure 1: The generated unigenes for both lily and tulip species and defining the orthologous sequences between them.

A Blast analysis using the non-redundant protein database (nr) from NCBI with an E value threshold of 1E-15 was performed using Blast2Go software (Conesa et al., 2005). Around 49% of lily unigenes (25,439 unigenes), around 30% of the tulip unigenes (24,748 unigenes), and around 88% of the orthologous sequences (9,584 unigenes) had at least one significant blast hit. As was expected, *Oryza sativa* (the most sequenced and annotated monocot species) showed to be the closest species to both lily and tulip because most first hits were with sequences from this species. Having only 49% and 30% of lily and tulip genes annotated respectively demonstrates the very rich source of not yet identified genes that need to be discovered and annotated which could indicate a unique position of bulbous plants. However, it is also possible that genes from lily and tulip deviate significantly at the sequence level from the existing orthologous genes in databases at the threshold value of 1E-15.

Gene ontology provides a structured and controlled terminology to describe gene products according to three categories: molecular function (refers to a biochemical activity of a gene product without stating where or when the event happens), biological process (refers to a biological objective to which the gene product contributes), and cell component (refers to the place in the cell where a gene product is active) (Ashburner et al., 2000). Since genes can be part of different pathways or have more than one function in the same time, the same gene can have more than one GO description (GO term) and thus belong to more than one of the earlier mentioned categories. The annotated unigenes of lily, tulip, and the orthologous sequences between lily and tulip were used to assess the GO term using Blast2Go (Conesa et al., 2005). The GO term of lily unigenes was divided into: 42% (molecular function), 31% (biological process), and 27% (cellular component). In tulip, GO term was divided into: 19% (molecular function), 42% (biological process), and 39% (cellular component) unigenes.

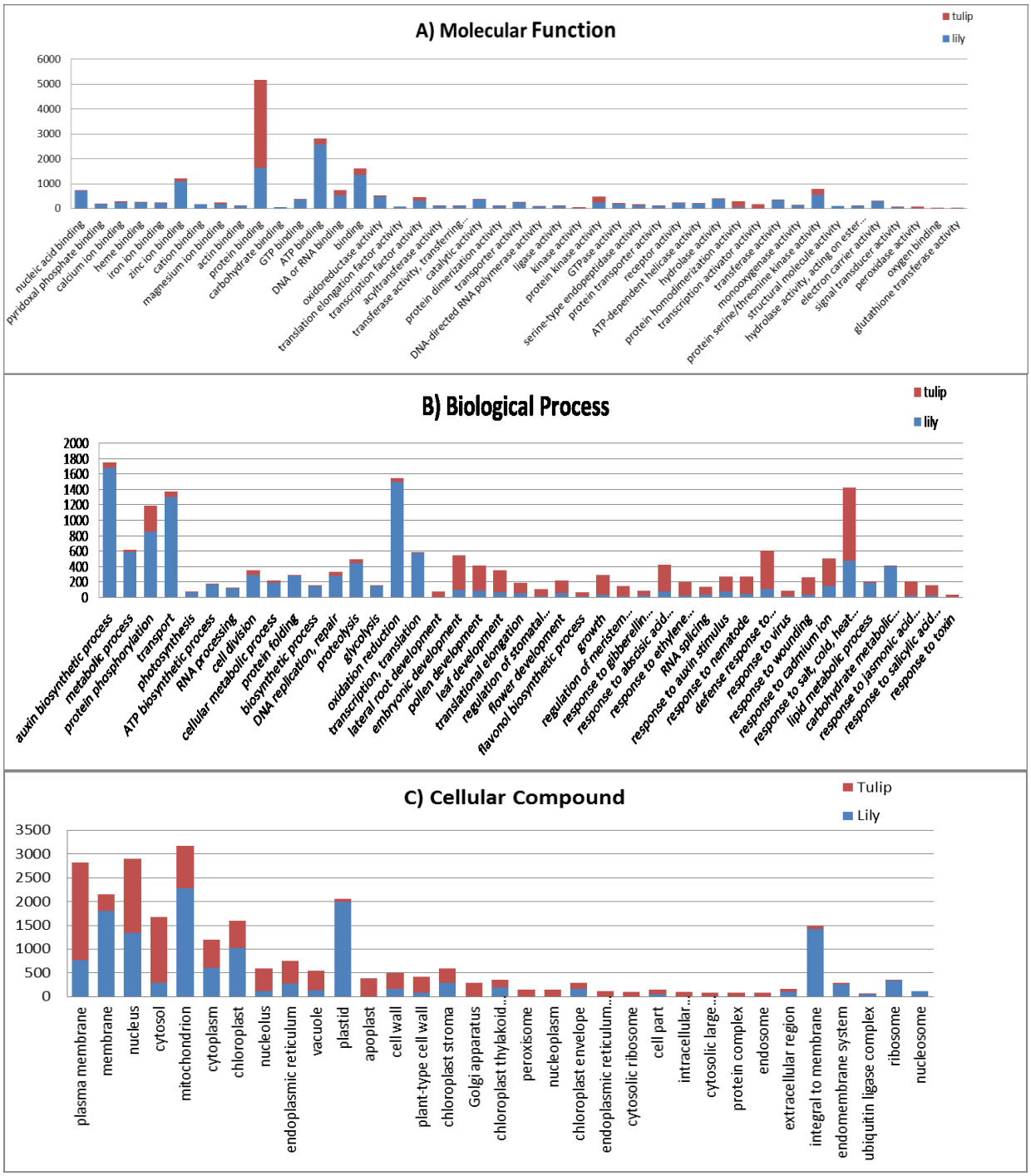


Figure 2: Representation of transcriptome ontology assignments for of lily and tulip unigenes from 454 sequencing data. **A**, the GO terms of molecular function, **B**, the GO terms of biological process and **C**, the GO terms of cellular compound category.

Both species showed to have similar GO terms in the three categories, which were also similar to rice (Liu et al., 2010). The differences were in the amount of unigenes annotated for each GO term. In molecular function category, the most represented GO terms were of binding function

such as 'protein binding', 'ATP binding', 'binding', 'nucleic acid binding' in addition to all types of activities such as 'protein kinase activity', 'transferase activity', 'transporter activity', 'catalytic activity' and 'oxidoreductase activity' (Fig. 2A). These GO terms together with many other GO terms that were identified in lily and tulip (Fig. 2A) were identified as well in *Medicago truncatula*, *Cucurbita pepo*, *Cucurbita melo*, and *Oryza sativa* (Blanca et al., 2011; Cheung et al., 2006; Gonzalez-Ibeas et al., 2007; Liu et al., 2010). Ion binding termenology such as 'calcium binding', 'iron binding', and 'zinc binding' were highly represented in lily (Fig. 2A), similar to the case in olive leaf (Ozdemir Ozgenturk et al., 2010).

In biological process category, there were clear differences between lily and tulip in the amount of unigenes assigned to each GO term (Fig. 2B). Lily's unigenes were more concentrated in activities like 'metabolic process', 'carbohydrate metabolic', 'lipid metabolic process', 'transport', 'oxidation reduction', 'protein phosphorylation', and 'ATP biosynthetic process' whereas response to biotic and biotic stresses such as responses to salt, heat, cold, nematode, bacteria, virus, and fungus stresses were more represented in tulip (Fig .2B). Responses to hormone's stimulus such as (ABA, gibberellin, jasmonic acid, salicylic acid, auxin, and ethylene) were presented in both species similar to rice (Liu et al., 2010).

The 'flower development', 'embryonic development', and 'pollen development' unigenes were more enriched in tulip compared with lily. This might be related to the fact that tulip growing phase (from leaves developing till seed formation) is short (7-12 weeks) compared with lily (5-6 months). Meaning, flowering and vegetative growing stages in tulip are integrated while they are separated in lily, consequently, flowering genes (control flower, pollen, and embryonic development genes) will be expressed in tulip together with other development process genes in this stage of young leaves. On the other hands, high level of 'auxin biosynthetic process' was recorded in lily, which might reflect the central on-going process which are mainly plant-cell elongation, apical dominance (inhibit growth of lateral buds), and rooting process which are all controlled by auxin.

The GO terms of cellular compound category showed significant representation of 'plasma membrane', 'membrane', 'nucleus', 'cytosol', 'mitochondrion', 'chloroplast', 'plastid', and 'integral to membrane' (Fig. 2C) which was similar to previous studies (Blanca et al., 2011; Gonzalez-Ibeas et al., 2007; Ma et al., 2006). All unigenes of mitochondria, chloroplast, and plastid that were defined here (Fig. 2C), are very interesting for phylogenetic studies while should be excluded when thinking to develop SSR or SNP markers for mapping studies.

The GO assessment of the 9,584 annotated orthologous unigene sequences divided into: 15% (molecular function), 49% (biological process), and 35% (cellular component). A summary description with the number of unigenes annotated in each GO category for the orthologous genes is provided in Figure 3. Genes essential for growing and defense processes are showed to be the main orthologous sequences between the two species. In biological process category, the

most frequent terms are ‘cellular process’ and ‘cellular component’ which are needed for growing and development processes. Genes involved in response to biotic, abiotic, and endogenous stimulus were also defined (Fig. 3B). Under molecular function category, mainly binding and catalytic activity were identified (Fig. 3C). Overall, majority of orthologous genes were housekeeping genes. More detailed data is available for all annotated lily and tulip unigenes and also for the orthologous sequences that will serve as a major resource for further research.

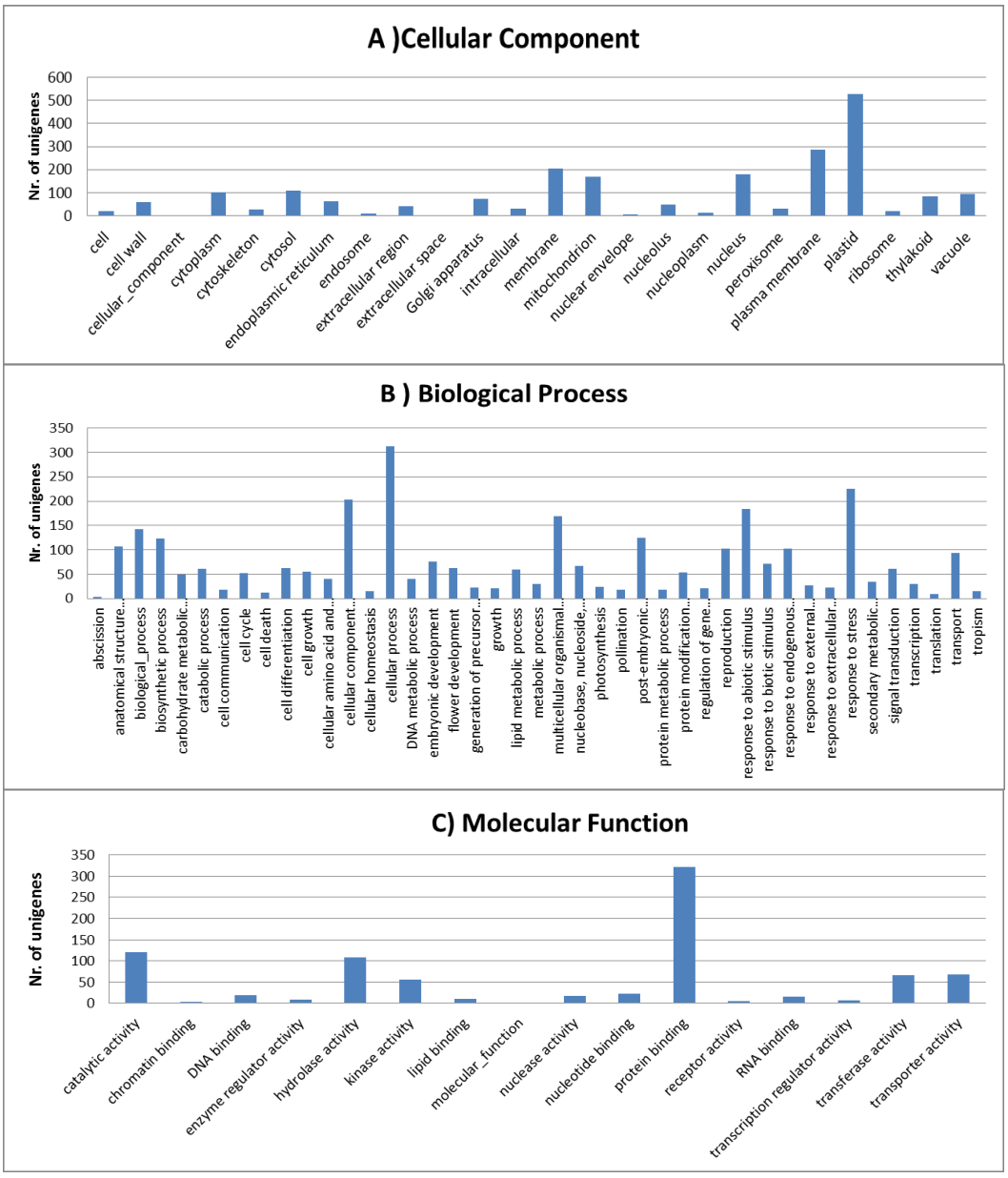


Figure 3: Representation of transcriptome ontology assignments for the orthologous sequences between lily and tulip from 454 sequencing data. **A**, the GO terms of the molecular function, **B**, the GO terms of the biological process and **C**, the GO terms of the cellular compound category.

Identification of common SNP and SSR markers within and between lily and tulip

Exchanging genetic information between two related species by linking their genetic maps would be of great interest. This linking will facilitate comparative mapping of genes across distantly related plant species by direct comparison of DNA sequences and map positions such as between wheat and barley, tomato and potato, and *Arabidopsis* and *Brassica* (Erpelding et al., 1996; Lan et al., 2000; Tanksley et al., 1992).

Identification of common markers based on orthologous sequences of both species will provide a set of common genetic loci that can be implemented for comparative mapping. For this, SNP and EST-SSR markers were developed from the parents of lily ('Connecticut King' and 'White Fox') and tulip ('Cantata' and 'Kees Nelis') mapping populations.

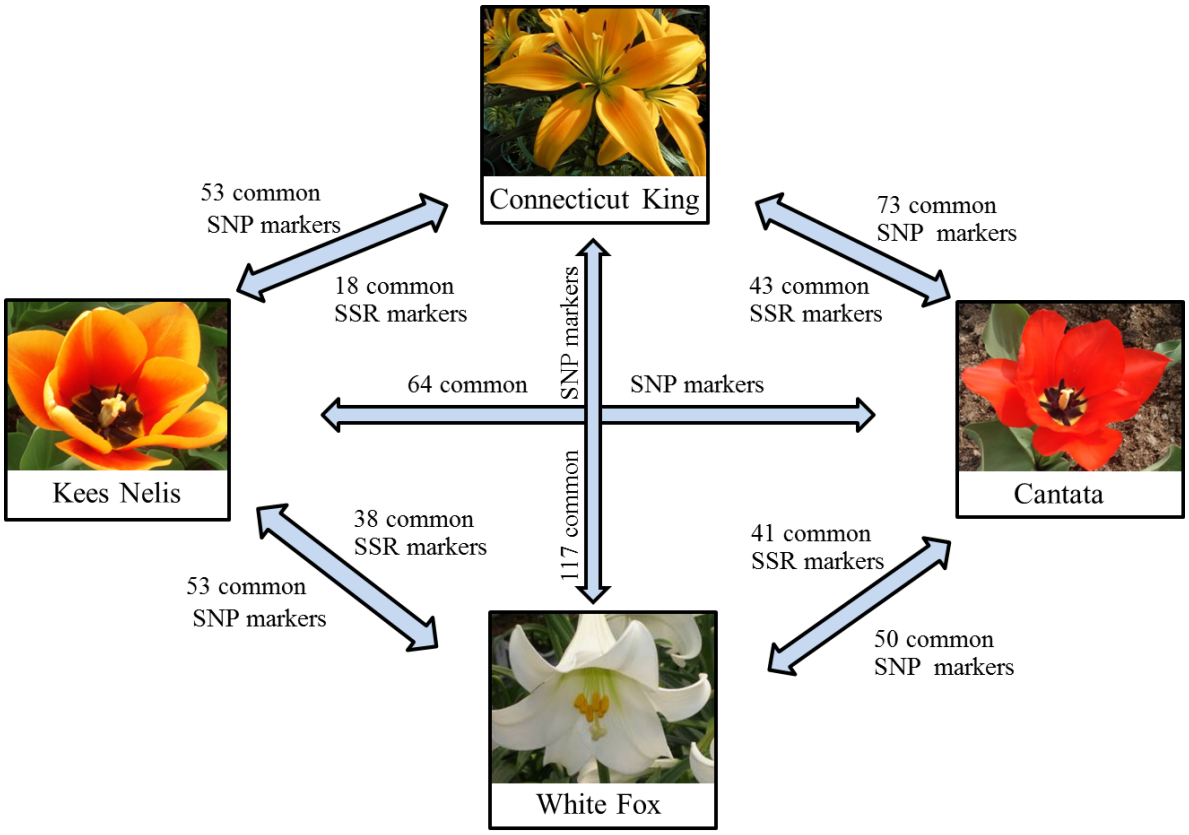


Figure 4: Common SNP and SSR markers identified among the four parents of the lily and tulip populations.

As a result, 'Connecticut King' showed to have 53 and 73 SNP markers in common with 'Kees Nelis' and 'Cantata', respectively; 'White Fox' has 53 and 50 common SNP markers with 'Kees Nelis' and 'Cantata', respectively (Fig. 4). As for common SSR markers, 'Connecticut King' showed to have 18 and 43 common EST-SSR primer pairs with 'Kees Nelis' and 'Cantata', respectively. Similarly, 'White Fox' has 38 and 41 common EST-SSR primer pairs with 'Kees Nelis' and 'Cantata', respectively. Thus, 229 common SNP and 140 common EST-SSR markers were identified between the lily and tulip populations. This is higher than the number of SNP and

EST-SSR markers identified by comparing ESTs containing SNP or EST-SSR markers of *Jatropha* with the ESTs of castor bean (215 common markers), poplar (202 common markers) and *Arabidopsis* (192 common markers) (BLAST X, 1E–5), which were generated for comparative genome analysis to identify conserved syntenies between *Jatropha* and castor bean, popular, and *Arabidopsis* (Wang et al., 2011). Also, common SNP markers between the parents of lily population and between the parents of tulip population were identified. 'Connecticut King' and 'White Fox' have 117 common SNP markers, and 'Cantata' and 'Kees Nelis' have 64 common SNP markers.

Using common markers between lily and tulip will enables running a comparative study based on the genetic maps of these two species. These common markers will be genotyped in the two populations that will allow us: to link the linkage groups of the two species, study the macro-synteny by comparing the order of common markers, and to transfer knowledge such as QTL regions between the two species. The efficiency of these markers in comparative study depends largely upon how many of these markers will be mapped on the genetic maps and also on how well these markers will be distributed over the chromosomes. This also will define if the current number of markers is sufficient to carry out such a study or that more markers should be generated.

Conclusion

The 454 pyro-sequencing provides a foundational transcriptomes resource for markers development and comparative genomics studies for species with an uncharacterized genome. We sequenced leaves transcriptomes of four lily and five tulip genotypes using 454 pyro-sequencing that was resulted in developing more than eight thousands unigenes for tulip, and five thousands for lily. Unigenes were used to identify SNPs markers suited for high throughput genotyping purposes, and to mine for SSR markers for lily and tulip. We identified molecular markers that are specific for each genotype, and also markers among genotypes of lily and tulip that can be used for genotyping a wide range of genotypes with different genetic background for association and genetic diversity studies. Additionally, generating these unigenes allowed the identification of orthologous genes (micro-seneteny) between the two species that were annotated and described according to Gene Ontology terminology. Common markers between lily and tulip were identified, which will allow running comparative genemoic studies by linking the linkage maps of the two species and look for macro-synteny. Overall, for species with little genetic resources available applying NGS technology opens the door for a wide range of genetic and genomic studies.

Chapter 5

Genotyping and Mapping of SNP Markers in *Lilium* L.

Arwa Shahin[1,2], Jaap M. van Tuyl[1], Richard G.F. Visser[1], Paul Arens[1]

[1]Wageningen UR Plant Breeding, Wageningen University and Research Centre, P.O. Box 386, 6700 AJ, Wageningen, the Netherlands

[2]Graduate School Experimental Plant Sciences, Wageningen University

Abstract

There is considerable interest in applying next generation sequencing and SNP genotyping techniques to accelerate genetic studies in different species. Genetic studies in *Lilium* have been challenging since this out-crossing species has a very complex and huge genome (36 Gb). In this study, we applied one of the new genotyping techniques (KASP, KBiosciences competitive Allele Specific PCR) to genotype SNP markers. These SNP markers were derived from 454 pyro-sequencing of a cDNA library constructed of cultivar 'Connecticut King'. A total of 225 SNP markers were used for genotyping two half-sib mapping populations of lily (LA and AA populations) using KASP technology. The genotyping success rate was 75.5% (170 SNP worked) from which 102 SNP markers were polymorphic (45%). A total of 94 (42%) and 85 (38%) SNP markers were mapped on the genetic maps of the LA and AA populations, respectively. This lead to an increase of 37% and 39% in the total number of mapped markers, and an increase of 48% and 32% in the total genome coverage in the LA and AA genetic maps with a marker density of one marker every 4 and 5 cM, respectively. Some SNP markers that deviated from the expected Mendelian ratio, or that had null alleles were re-scored manually. SNP markers showed to be very efficient in mapping in both populations and were distributed over chromosomes. In conclusion, development of SNP markers of transcriptome sequences generated using next generation sequencing technology of uncharacterized lily genome was successful, and SNP markers showed a high efficiency in mapping.

Introduction

The vast development in high-throughput sequencing technologies together with the fact the plant and animal genomes are packed with single nucleotide polymorphisms (SNPs), and the ability to genotype large numbers of SNPs over large sets of individuals have led to a revolution in their use as molecular markers (Lepoittevin et al., 2010; Pavy et al., 2008). The SNP markers have been developed and genotyped using one of the next generation sequencing technologies together with one of the high throughput genotyping methods in several species such as *Zea mays*, *Cucurbita*, *Eucalyptus*, and *Pinus* (Blanca et al., 2011; Chancerel et al., 2011; Grattapaglia et al., 2011; Yan et al., 2010). However, to the best of our knowledge, these high throughput new sequencing and genotyping technologies were not applied so far on any ornamental flower crop.

Genus *Lilium* L. includes around 100 species and 9,700 cultivars (International Lily register, http://www.lilyregister.com/). The most important cultivated groups are: Longiflorum (L), Trumpet (T), Asiatic (A), and Oriental (O) hybrid groups. These hybrids contain important traits such as resistance to *Fusarium oxysporum*, *Botrytis elliptica*, and viruses in addition to commercial characteristics. Inter-specific crossing between hybrids of different groups is performed to combine traits of interest to further improve the crop. However, this puts additional

requirements on selection process given that several QTL traits from different backgrounds should be combined.

Molecular linkage maps are important for genetic studies and are necessary for quantitative and qualitative trait analysis that facilitate molecular assisted breeding (Dubey et al., 2011). However, despite the economic importance of lily, only very few genetic mapping studies have been performed. The most recent genetic maps have been published by Shahin et al. (2011). These maps showed an improvement in the number of mapped molecular markers, genome coverage (measured by cM), and marker density (average distance between markers in cM) compared with previous studies (Abe et al., 2002; Van Heusden et al., 2002). Shahin et al. (2011) developed two genetic maps for cultivar 'Connecticut King' that was used as parent in two lily crosses: LA and AA populations. The LA genetic maps were constructed by using 411 DArT (Diversity Array Technology) and NBS (Nucleotide Binding Site) molecular markers. The AA genetic maps were constructed by using 295 AFLP (Amplified Fragments Length Polymorphism), DArT, and NBS molecular markers. Genetic maps of 'Connecticut King' covered 1,642 and 1,539 cM with a marker density of one marker every 4 cM and 5 cM, for the LA and AA populations respectively.

For molecular marker development and linkage map saturation, SNP makers have become the markers of choice (Pavy et al., 2008). Abundance of SNPs throughout the genome, co-dominant scoring of SNP markers, ease of genotyping, and direct application of SNP markers without the need for further marker conversion all together support this choice. This study is a continuation of the work in Chapter 4 concerning developing SNP markers of lily for genetic mapping and QTL identification purposes. The aims of this study were: 1) verification of SNP markers generated from a cDNA library of 'Connecticut King' using 454 pyro-sequencing technology, 2) mapping SNP makers on the genetic linkage maps of two lily populations and study SNP mapping efficiency, and 3) look for improvements in genome coverage and marker density in the maps.

Material and methods

Plant Material

For genotyping, two mapping populations were used. The first is a F1 population of 98 genotypes (LA population: *L. longiflorum* 'White Fox' x Asiatic 'Connecticut King') that was produced in 2000 using cut style pollination and embryo rescue. The second is a BC1 AA population of 98 individuals (Straathof et al., 1996; Van Heusden et al., 2002) that was produced in 1989. It is a backcross of 'Connecticut King' with 'Orlito' (= 'Connecticut King' x 'Pirate'). 'Connecticut King' is a common parent in the two populations and it is a well-known Asiatic cultivar resistant to LMoV (Lily Mottle virus) and resistant to *Fusarium oxysporum* (Shahin et al., 2011).

SNP marker development

SNP marker development included the following steps: RNA isolation of the leaves of 'Connecticut King, cDNA library construction and normalisation, 454 pyro-sequencing, sequence assembly, and SNP marker detection are explained in Chapter 4. As a result, 406 SNP markers (flanked by 50 bp on either side free of SNPs) were developed for 'Connecticut King' and were available for genotyping (Chapter 4). Randomly, 225 out of the 406 SNP markers were used for genotyping. SNP markers were named using the abbreviation (SNP), followed by a number referring to the contig that SNP was generated from, *e.g.*: SNP_78 refers to SNP marker that was generated from contig 78.

SNP genotyping

Genotyping was done by KBioscience Ltd (http://www.kbioscience.co.uk/) using fluorescence-based competitive allele specific PCR (KASPar) assay. Young leaves of the parents of the two mapping populations: 'Connecticut King', 'White Fox', 'Orlito', and grandparent 'Pirate', in addition to the progenies of AA and LA populations were sampled and send to KBioscience for genotyping. Water was used as a control and parents were included in duplicate to test reproducibility. SNP data were visualized using SNPviewer2 software (http://www.kbioscience.co.uk/software/snpviewer/snpviewer_help/index.htm). The expected segregation ratios for co-dominant SNP markers were 1:2:1 for SNPs that were heterozygous in both parents or 1:1 for SNPs that were heterozygous in only one parent respectively. Segregation patterns of all SNPs were checked, and those SNPs that showed strange segregation of the homozygous and heterozygous groups were re-scored manually (Fig. 1).

Genetic mapping

The SNP marker data, together with already available NBS and DArT markers (Shahin et al., 2011), were used to re-construct linkage groups for LA population. Similarly, these SNP markers were joined with the already available AFLP, NBS, and DArT markers to re-construct the genetic map in AA population. Genetic maps of both populations were constructed for the 'Connecticut King' only, since all the available markers were developed for this parent (it is the parent that has the resistances to *Fusarium oxysporum* and lily mottle virus).

Genetic maps were constructed using JoinMap 4.1 (Van Ooijen, 2006). Both crosses were analysed using the "Cross Pollination" option even though AA population is a BC1. This is beacues in outcrossing heterozygous species, a BC1 cross can be considered as an F1 cross. The segregation ratio of alleles for each locus was evaluated by Chi-square testing with a significance threshold of P=0.05. Recombination frequencies were converted into map distances in centi Morgans (cM) using Haldane's mapping function. Grouping was based on the independence LOD (Logarithm of odds) parameter, using regression method. All markers were first grouped at a minimum LOD threshold of 3.0. MapChart (Voorrips, 2002) was used to draw the genetic linkage maps.

Results

SNP marker genotyping

The 225 SNP markers were used for genotyping progeny of LA and AA populations using KASP technology, (KBioscience: http://www.kbioscience.co.uk/). Out of the 225 SNP markers, 170 SNP markers genotyped successfully (75.5%), 68 of them were monomorphic, and 102 were polymorphic SNPs amounting to a conversion rate of 45% (Table 1).

Table 1: The SNP markers genotyping results for LA and AA populations

	LA population	AA population
Genotyped SNP	225	225
Successfully genotyped	170 (75.5%)	170 (75.5%)
Polymorphic SNP	102 (45%)	102 (45%)
Deleted SNP due to strange segregation	3	5
Used for mapping	99	97
Mapped SNP	94 (42%)	85 (38%)

Segregations of SNP markers were of 1:1 or 1:2:1 type depending on polymorphism in parents. If only 'Connecticut King' is polymorphic then a 1:1 segregation ratio is expected, while a 1:2:1 ratio is expected if the other parent is also polymorphic. Visualizing segregation of SNPs using SNPViewer2 allowed us to check the segregation of the SNPs. In case of <ABxAB> marker type, three clusters: AA homozygote, BB homozygote, and AB heterozygote are expected, while in case of <ABxBB> or <AAxAB> marker types two groups: homozygote cluster (AA or BB homozygote), and AB heterozygote cluster are expected. Some SNP markers showed strange segregation (not fitting with Mendelian segregation): 6 SNP markers in LA and 17 in AA populations (see Fig. 1 A, B for examples). For instance, SNP_17715 marker is of <ABxAB>, and three clusters are expected, however there was no BB cluster (Fig. 1A). Segregation of this marker could not be explained and subsequently such markers were skipped form mapping. In another example, genotyping LA population using SNP_10602 (<ABxAB> type) resulted in two clusters that were scored as AB and one AA cluster, whereas BB cluster was missing (Fig. 1B). By changing the scoring of the genotypes of one of the two AB clusters to BB, this marker could be used for mapping (Fig. 1B). As a result of re-scoring the deviated markers, 3 markers of LA and 12 markers of AA populations were corrected and could be mapped successfully.

Few other SNP markers, 7 in LA and 2 in AA populations showed strange segregation due to the presence of null allele (ø). In LA population, two null alleles (øø) can be inferred for a certain marker when 'White Fox' shows no call in both replicate samples. Consequently, genotyping the progeny with such marker <øøxAB> will result in two clusters (Aø and Bø) that will be visualized as (AA and BB) in the SNPViewer2 (Fig. 1C) since the genotyping technology cannot distinguish between (AA and Aø) allelic combinations. Such markers can still be used in mapping by converting them into a bi-allelic single parent marker <AAxAB> type and change the scoring of the progeny of cluster (Bø) into AB (Fig. 1C). Doing so, we could map all SNP

markers that showed homozygote null allele (øø) in 'White Fox'. It is also possible that only one allele of 'White Fox' is a null allele ('White Fox' is Aø or Bø). Such markers <AøxAB> will result in four allelic combinations with equal segregation ratios (AA, Aø, AB, and Bø). This, however, will be visualized as only three groups since AA and Aø will be presented as one AA cluster. Consequently, the segregation ratio of such marker will be (2:1:1) instead of (1:1:1:1). This case was not recorded in our data for LA population.

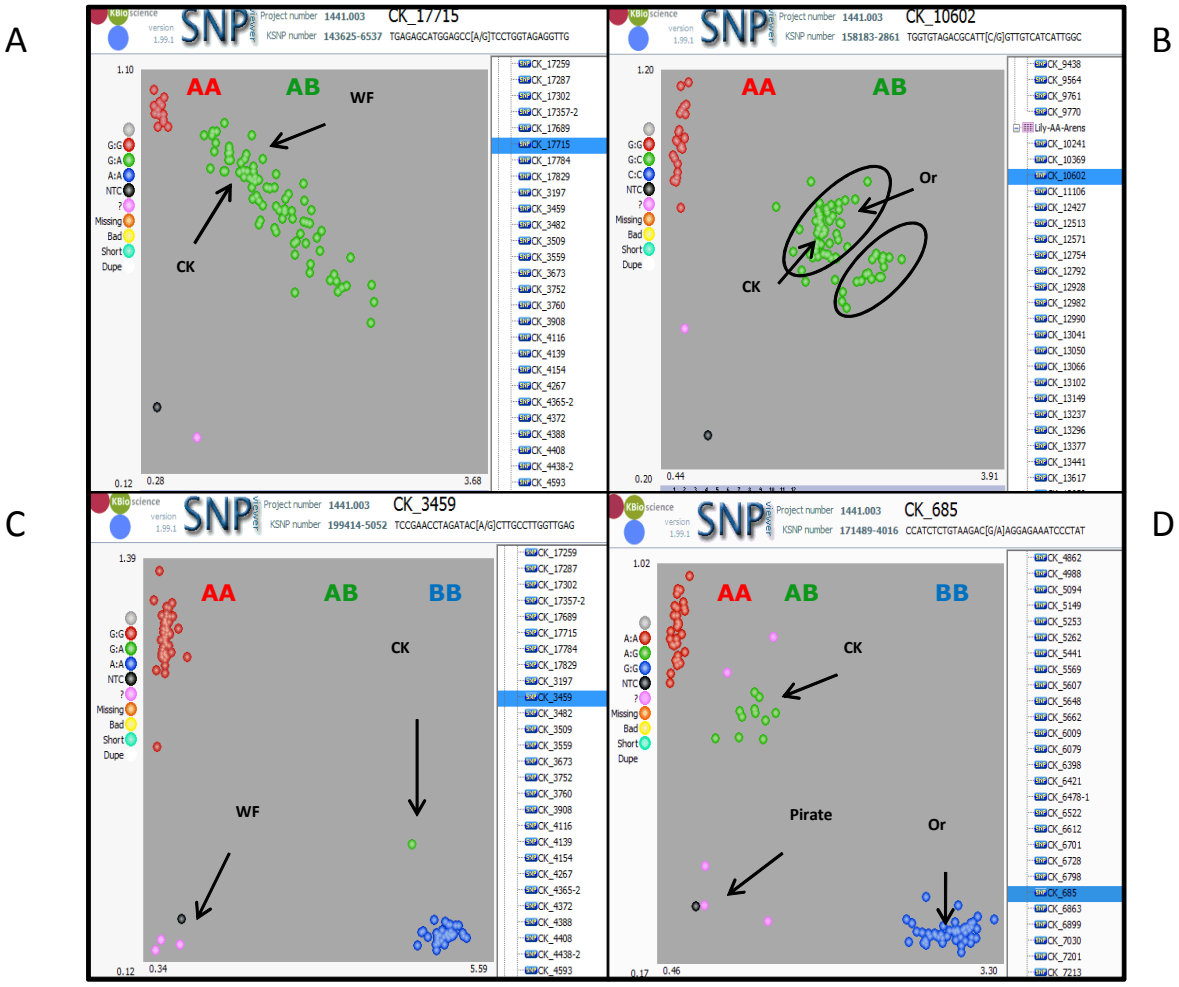


Figure 1: Scoring of SNP markers genotyped using KASP technology, and visualized using SNPviewer2 software. **A**, **B**: odd segregation of <ABxAB > marker type in the LA and AA populations respectively. **C**, **D**: segregation of SNP markers in case of null allele in the LA and AA populations respectively. CK refers to 'Connecticut King', Or refers to 'Orlito', WF refers to 'White Fox'. Red dots refers to the genotypes that have AA alleles, blue dots refers to the genotypes have the BB alleles, and the green dots refers to the genotypes have AB alleles for certain SNP.

In AA population, a null allele (ø) can be identified by genotyping the grandparent 'Pirate' that will show no call, while the father 'Orlito' will be either Aø or Bø (Fig. 1D). Meaning, in case of a null allele present in AA population genotyping a marker type <ABxAø> will result in four allelic combinations with equal segregation ratios (AA, Aø, AB, and Bø), which will be visualized as three clusters with skewed segregation (2:1:1) as was explained above. It is also

possible that 'Pirate' might have only one null allele (Aø), consequently 'Orlito' will either have no null allele (AA, and marker type <ABxAA> that segregates normally), or have one null allele (Aø, and a marker type <ABxAø> that segregates as what mentioned above).

These markers <ABxAø> can still be used for mapping by considering them as bi-allelic markers that are fully informative from mother side (by calling them <ABxAA>, consider AA and Aø as one AA cluster and combine AB and Bø into one AB cluster). From the father side <ABxAø> only the allelic combinations AB and Bø are informative since we can deduce which allele has been passed on; whereas, both AA and Aø will be not informative since we cannot distinguish if the allele passed on from mother or father, thus AA and Aø can be changed into missing data. In this study, we converted the two makers that showed to have a null allele in AA population into bi-allelic markers that segregate from mother 'Connecticut King' since these markers will be fully informative. Doing so, we could map all SNP markers that showed null allele.

As a result, 102 polymorphic SNP markers were identified in lily that represents a conversion rate of 45 % (the number of polymorphic SNPs divided by total number of SNP markers used for genotyping). Next, SNP markers that showed strange segregation and could not be re-scored manually were excluded and therefore 99 and 97 SNP markers in LA and AA populations respectively were used for genotyping (Table 2). List of SNP markers that are mapped on both populations and their annotation results (top hit) from blasting to NCBI (BlastX) were included in Table 2.

Table 2: Mapping information of LA and AA populations

	LA population				**AA population**			
	Total	Previous maps	New maps	Mapped markers %	Total	Previous maps	New maps	Mapped markers %
AFLP	--	--	--	--	301	134	154	51
DArT	552	380	441	80	244	91	96	39
NBS	34	31	30	88	155	70	74	48
SNP	99	--	94	95	97	--	85	89
Total	685	411	565	82.5	797	295	409	51
Coverage cM		1,642	2,438			1,539	2,035	
Marker/cM		4.1	4.3			4.9	5.2	

Construction genetic linkage maps

LA population

A total of 99 polymorphic SNP markers together with 34 NBS and 552 DArT markers that were already available (Shahin et al., 2011) were used to re-construct genetic map for LA population. Linkage groups for 'Connecticut King' were constructed with a LOD >4. A total of 565 markers (94 SNP, 30 NBS, and 441 DArT markers) were mapped on 22 LA linkage groups (LGs). The mean Chi-square of linkage groups ranged between 0.06 and 1.3. The LA linkage maps span

2,438 cM with an average marker density around 4 cM. Three regions showed skewed segregation (*P*=0.05) clustered on linkage groups: 1b, 14, and18.

AA population

Linkage groups of 'Connecticut King' were re-constructed for AA population using 97 polymorphic SNP markers together with 155 NBS (53 <ABxAA> and 102 <ABxAB>), 244 DArT (62 <ABxAA> and 182 <ABxAB>), and 301 AFLP (183 <ABxAA> and 118 <ABxAB>) markers that were available previously (Shahin et al., 2011). Linkage groups of this population were mapped with a LOD > 4. More than 50% of the used markers for mapping were of <ABxAB> type that was scored dominantly (1:3). This marker type has low information content as only 25% (absence of band) of their data can be used for mapping. Consequently, several linkage groups showed suspect linkage and thus difficulties in mapping. To solve this challenge, linkage groups of LA population were used as references for mapping in AA population, using mainly SNP markers (their presence and their order on LGs) to correct the mapping of AA LGs that showed suspect linkage. This was done by comparing mapping positions of the common SNP markers in both populations. Accordingly, if few SNP markers are mapped on one LA linkage group and on one AA LG despite the fact that this AA LG showed to have suspect linkage, we kept it as one linkage group and we discarded (1:3) markers. While if these SNP markers were mapped on two LA LGs, we split up the AA linkage group (that have suspect linkage) into two LGs. In all cases, discarding the (1:3) markers improved the mapping in this population. Additionally, having SNP markers in common between the two populations helped in building and structuring the linkage maps of AA population. This AA map comprises 22 linkage groups of which 5 small LGs with 3-5 markers. The mean Chi-square of the linkage groups ranged between 0.4 and 1.9. The AA genetic maps spans 2,438 cM with an average density of a marker each 5 cM. Three regions showed skewed segregation (*P*=0.05) which clustered on linkage groups: 3, 12b, and 17.

Map alignment

The current genetic maps of LA and AA populations were compared with previously published genetic maps for those two populations (Shahin et al., 2011). Majority of current linkage groups were similar to the old maps except for some groups that were linked to each other, *i.e.* LG9 joined LG1 (LA9 now is LA1b, and AA9 is part of AA1b in this maps), LG13 joined LG2 (LA13 was integrated in LA2 and AA13 integrated in AA2a), and LG18 joined LG7 (AA18 now is AA7a) in current maps (Fig 2).

Previously, a comparison between genetic maps of two populations was made using a few common DArT and NBS markers (21 DArT and 8 NBS markers, Shahin et al., 2011) that enabled the alignment of 15 LGs. Now, with the use of SNP markers, all LGs of the two populations could be aligned and the previous alignments of the LGs from Shahin et al. (2011) were fully confirmed.

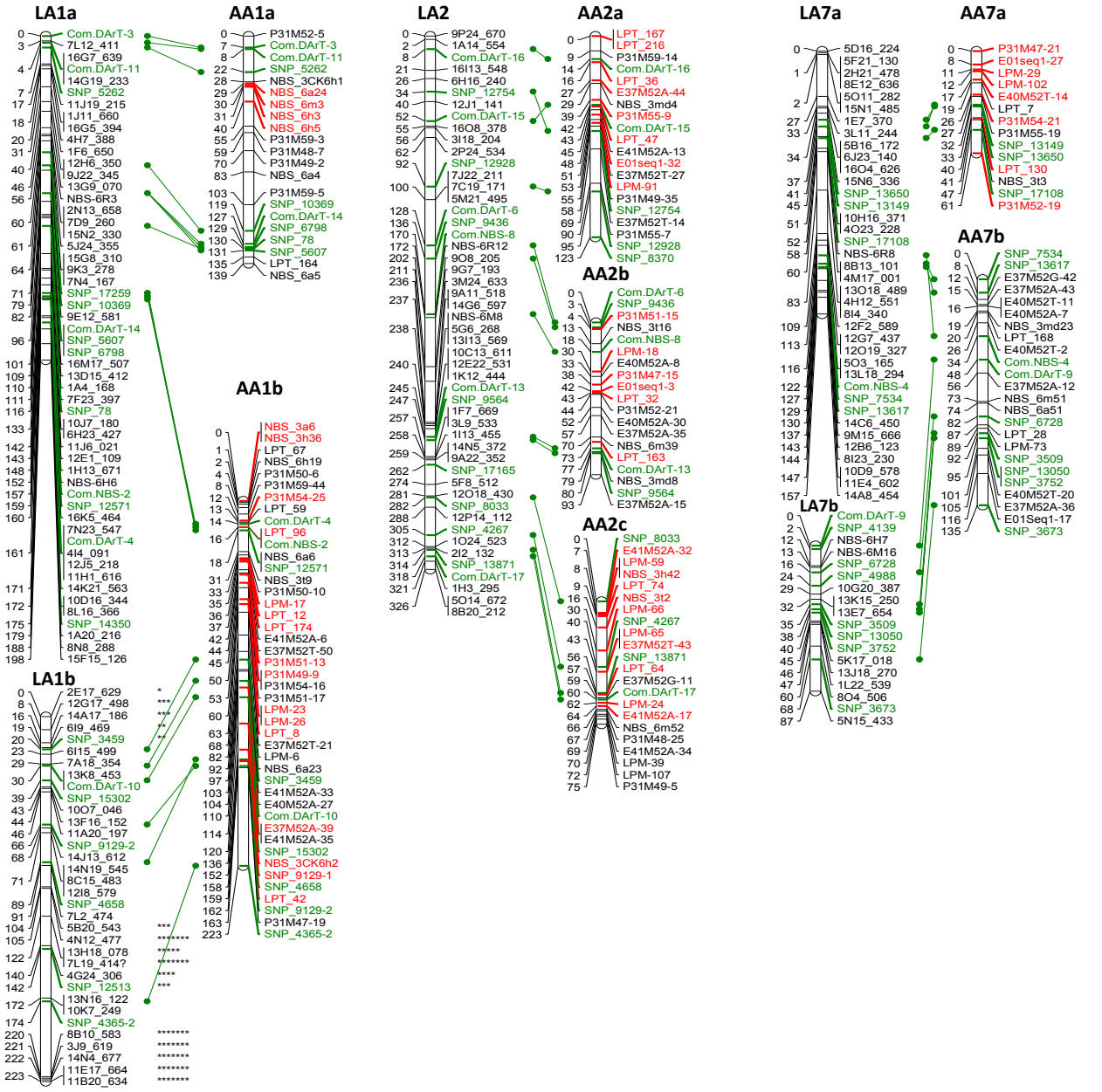


Figure 2: The linkage groups 1, 2, and 7 of LA and AA populations aligned according to the common DArT, NBS, and SNP markers. All SNP markers and the common DArT and NBS markers are in green. The <ABxAB> markers are in red. The skewed regions are shown by stars.

The order of common markers (SNP, DArT, and NBS) between the two populations was the same in majority of linkage groups. However, some differences in the order were observed, *e.g.* SNP_9129-2 and SNP_4658 on the bottom of the LA1b and AA1b showed to have different order. This might be due to the presence of the (1:3) markers in AA population that cannot be mapped accurately and thus may cause a change in markers order.

Current maps of LA and AA populations were compared with previously published maps in the sense of number of mapped markers, the coverage, and marker density (Table 2). A total of 565 markers were mapped on LA maps compared with 411 markers mapped on the previous map. The increase was due to SNP markers (94), in addition to 61 DArT markers that were not mapped previously. This increase in number of mapped markers caused growth in the coverage of lily genome to reach 2,438 cM (increase of 796 cM) compared with the old map, with the same marker density (one marker per 4 cM, Table 2). Similarly, 114 new markers mapped on current AA genetic maps (85 SNP, 5 AFLP, 20 DArT, and 4 NBS markers) which caused an increase of around 500 cM in genome coverage, at the same marker density (one marker per 5 cM) compared with the old AA maps.

Mapping efficiency of marker types

There were differences in mapping efficiency of different molecular marker types used for mapping in both populations. In LA population, both NBS and SNP markers showed high efficiency with 88% and 95 % of NBS and SNP markers being mapped, while DArT markers were slightly less efficient with 80% being mapped (Table 2). Overall, 82.5% of markers genotyped in LA populations were mapped. In AA population, the overall efficiency of markers was low (51%). In detail, SNP markers were highly efficient (89%), while AFLP, DArT, and NBS markers showed an efficiency of 50% or less. This is mainly due to the dominant scoring of <ABxAB> markers (which were the majority in this back cross population) that were subsequently excluded since they tend to cause more tension in the maps (Shahin et al., 2011). Mapping efficiency of SNP markers in AA population was 89% which is comparable to 95% in LA population.

Discussion

SNP marker genotyping

In this study, we could genotype for the first time SNP markers in flower bulb crops. We confirmed the usability of next generation sequencing technology in lily and the effectiveness of sequence assembly and SNP marker identification steps that were explained in Chapter 3. A total of 225 SNP markers of 'Connecticut King' were tested, from which 170 markers worked (75.5%). The genotyping success rate (75.5%) was comparable to previous studies in outcrossing species such as: maritime pine (23.8 GB) 66.9% in which a combination of *in silico* and *in vitro* SNPs were used (Lepoittevin et al., 2010), 63.6% also in maritime pine *in silico* SNP (Chancerel et al., 2011), and less than 92% in maize in which all the SNPs were BlastN (1e-12) against the maize sequence in gene bank and only the top blast-hits were used for genotyping (Yan et al., 2010). The conversion rates (number of polymorphic SNPs divided by the total number of SNP markers used for genotyping) of 45% were higher than conversion rate for maritime pine (19.5 and 12.5% for two populations used) (Chancerel et al., 2011), close to the 42.5% conversion rate for pine *in silico* (Lepoittevin et al., 2010), and less than conversion rate of 69.2% of *in vitro* SNP in white spruce (Pavy et al., 2008). SNP markers that deviated from the expected Mendelian ratio

were also described in maize (Yan et al., 2010). Re-scoring these SNP markers showed to be effective since all the re-scored markers were mapped; however, it is a time consuming process. Thus in case of genotyping large number of SNPs excluding such markers will not be problematic.

Mapping of SNP markers

Developing genetic maps in lily is quite challenging due to two main reasons: the very large genome of lily (36 Gb, more than 280 times larger than *Arabidopsis* genome), and the very high chiasmata frequency (54.8 chiasmata were determined per complete set of diplotene bivalents in *L. longiflorum*, Stack et al., 1989). This means the need of a large number of markers to cover the genome and prevent linkage groups from splitting into two or more small linkage groups due to insufficient markers to bridge recombination hot spots. Using next generation sequencing technology to sequence the whole lily genome for marker development would be difficult due to the assembly problems in heterozygous organisms. Furthermore, lily genome has an abundance of repetitive elements which will make sequence assembly and thus SNP retrieval very challenging. Thus, developing markers using transcriptomes showed to be good option for such huge genomes *e.g.* white spruce, pine, and *Eucalyptus* (Chancerel et al., 2011; Grattapaglia et al., 2011; Pavy et al., 2008).

To increase mapping efficiency of markers in a backcross population (such as AA population in this case), co-dominate markers such as SNP or SSR are highly appreciated since more than half of the markers of such population will be of <ABxAB> type. Whereas for dominant <ABxAB> type markers only 25% of the data (3:1) is actually used for mapping (AA; *i.e.* the plants without a band), this percentage is doubled in case of scoring co-dominant markers (1:2:1) since both AA and BB can be distinguished and are informative situation for the parental segregation.

The SNP markers developed for the two populations improved old maps: the number of mapped markers increased 37% (154 markers) in LA population and 39% (114 markers) in AA population, the genome coverage increased 48% in LA and 32% in AA population, the marker density remained the same, the number of linkage groups decreased into 22 aligned linkage groups including even the very small linkage groups (3 markers). Decreasing the number of linkage groups were mainly due to mapping SNP markers that provided the link between small and large linkage groups, which indicated that SNP markers targeted genomic regions that were not targeted before with other type of molecular markers developed for lily. This might also explain the 48% increase in the genome coverage. Blasting the SNP marker's contigs (BlastX) *vs.* NCBI gene bank (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) provided an overview about the genes targeted by our SNP markers. Majority of the markers were predicting proteins and some have no blast hit (Table 2). Some housekeeping genes like tubulin and ubiquitin were mapped (SNP_17165 mapped on LA2, SNP_3752 mapped on LA7b, respectively) that can be used for further research to identify a reference gene for q-PCR. Additionally, SNPs provided common markers that were mapped in two different crosses. Common markers allowed combining genetic

information from two crosses such as resistance to *Fusarium* and LMoV, which can speed up marker assisted breeding (Shahin et al., 2011). Mapping and comparing the same trait in two populations allows the confirmation and validation of the identified QTLs.

Common SNP allow the construction of a consensus map of two crosses similar to studies in pine and white spruce (Chancerel et al., 2011; Pavy et al., 2008). However, in this study we did not construct a consensus map. For QTL mapping studies, mapping of two populations separately is preferred since the position of markers can be more reliably estimated and also the phase/allele of the markers can be followed easily.

Another benefit of common SNP markers between the two populations is to compare the synteny between the aligned linkage groups. The order of SNP markers on linkage groups should be the same in the two populations. Generally speaking, any difference in marker order between maps should be due to mistakes in mapping or due to possible biological phenomena such as paralogs or inversion events. In this study, we constructed the maps only for 'Connecticut King' thus the main reason for discrepancy in marker order that was recorded in a few cases (five) will be due to lack of markers in certain region (large gaps) or due to the (3:1) segregating markers that have less mapping power and thus might cause some changes in marker orders.

Genome size of lily was estimated to be 2740 cM using the hypothesis that one chiasma corresponds to 50 cM (Rodionov et al., 2002; Shahin et al., 2011). The average number of chiasmata per complete set of diplotene bivalents was determined in *L. longiflorum* at 54.8 (Stack et al. 1989). Thus the current maps of LA and AA populations cover 89% and 74% of lily genome, respectively. To estimate the number of markers needed to cover 95% of the expected *Lilium* L. genome (2740 cM), we used the formula developed by Lange and Boehnke (1982) $n = [\log(1-p)/\log(1-\frac{2c}{k}]$ where n is the minimum number of randomly distributed markers needed to cover a proportion of (p=95%) of a genome size k, at a maximum distance of c between two adjacent markers. A total of 204 markers are needed to cover the lily genome (k=2740 cM), assuming a random distribution of markers with a distance between markers of 20 cM, while 1,024 and 2,051 markers are needed to cover 95% of lily genome with a distance of 4 cM or 2 cM between adjacent markers respectively, which is more than the 373 and 935 markers needed to cover the 95 % estimated genome size of maritime pine with a maximum distance of 20 or 4 cM between markers, respectively (2500 cM; Chancerel et al., 2011). Using the same formula and assumption of randomly distribution of markers we could conclude that with a distance of 8 cM between markers, 511 markers are sufficient to cover 95% of lily genome. In LA population, 565 mapped markers cover around 90% of *Lilium* L. genome which is close to number calculated using the formula of (Lange and Boehnke, 1982) indicating that these markers are very close to the random distribution. Nevertheless, gaps were recorded in the linkage maps such as a gap of 60 cM at the bottom of LA 1b, indicating that some regions were under-represented compared with other regions. The availability of much more SNP markers developed for 'Connecticut

King' and 'White Fox' (1,171 and 1,292 SNP markers respectively, Chapter 4) allow us to genotype and map more SNP markers that can help to cover the entire lily genome at reasonable marker density and identify closely linked markers to important traits.

Conclusions

The present work validated a subset of SNP makers developed for lily and showed the usability of this type of markers to improve the genetic maps of complex and uncharacterized genomes like lily. Present genetic maps are the most advanced genetic maps for this species that will be enriched with more SNP markers and consequently improve QTLs resolution. Having two half sib populations provided a very good reference for mapping in which common markers were very useful in linkage group alignments and mapping position confirmation. Any discrepancy in mapping will draw attention to possible mistakes. Also, having two populations allows comparing the genetic information such as QTLs and transfer linked markers from one population to another. More importantly, QTLs can be validated when mapping same traits in two populations. With this, QTLs will be confirmed and molecular assisted breeding can be speeded up.

Table 2: List of SNP markers, their sequences and annotation (top blast-hit, from Blastx and their E values), that were mapped on LA and AA populations

SNP ID	SNP sequence	Description	E value
SNP_78	ATCATCGTCCAAGCAAGTCGAAGGATTCTAGTTCAAGCAC[A/G]CCCAAGGGTCCAAAAGCAACTAGGAAGGAATCCGAGACAT	no blast hit	
SNP_15507	CTCTTTCACCAGCATGCTGAACATAAGCATGATTCCCTGC[G/A]GGAACAGCCAACACAGTTGAGAAAATGAAACTCAAGATAA	Os03g0185600 [Oryza sativa Japonica Group]	4.00E-46
SNP_9438	TCTTCTCGCGACCTCGGGCTTTGAGTCGCTGCTCCTCATC[G/A]CGTCTTCTCTTTACTCTGATGGCTTGCCCTGTGAGAACAG	Os03g0185600 [Oryza sativa Group] >gb|ABF94356.1| expressed protein [Oryza sativa Group]	4.00E-46
SNP_11106	ATCGACTGCAGCTAGAATTCCTGGTGACCTCGAGGCCAAT[A/C]AAACAAATTCGACTTCATCCACTTGAAACTCTTTGTAGTT	30S ribosomal protein S17, putative [Ricinus communis]	1.00E-05
SNP_6728	AAGGGATGATGTGGAAGAAGCTGTGGGAGAAGATTCTGCT[A/C]TTTTGCACATCACCACACCCGGCCAGTTAACAGCTCTTAT	PREDICTED: hypothetical protein [Vitis vinifera]	1.00E-31
SNP_13871	TGCTGTATGTGACACCATTCCAGGCACTGGCAGTTCTTGC[A/G]GGGTTCTATATGATGAGGCATCCGAGGTTCCGTCACAGGA	PREDICTED: hypothetical protein [Vitis vinifera]	2.00E-34
SNP_15302	CTTCCGGCACATTACTGCTGATAATATCAGATTCTTCCAC[G/A]CCCAAGCTGGGGATCACAAACCCATCAGTATCCCAATCGT	hypothetical protein LOC100527167 [Glycine max]	2.00E-11
SNP_14781	TTGATCTTGGAGAGCTCGATGCAGCGATCGCGATCGTTGC[A/G]GACATGGAGACGGCCGGAATCCAAGTTCCTGATGAAACCC	unnamed protein product [Vitis vinifera]	1.00E-22
SNP_17259	CGTAGCATTGACAGCTTTGCATGCTCATTTAACTGGGCCA[A/G]TAGAGGCATCAATGCACCAGTACCAAGGACCAGATCTCGG	hypothetical protein OsI_01208 [Oryza sativa Group]	2.00E-113
SNP_7030	AATGCCAGATAACAGATATAGTTTTTCCTACAGCACATTC[C/A]GCATATTTTGCAGAGTACTGGTCACATTCTTTGAGTGCCT	PREDICTED: hypothetical protein [Vitis vinifera]	7.00E-29
SNP_17021	TCCGAGCATGCATGTGAGGCCACATTCGGCCGTTGCTCTC[A/G]TACGACGGCACATACACTTTTAAAACCTGGTTTCGCGTGA	hypothetical protein OsI_24908 [Oryza sativa Group]	2.00E-48
SNP_3459	CAAAAGGTGTTAAGCCAACATTATCTCCGAACCTAGATAC[A/G]CTTGCCTTGGTTGAGGCAGCTGGAGTAAAACGACTTCTCT	OSJNBb0026L04.9 [Oryza sativa Japonica Group]	2.00E-65
SNP_6079	TCGGTGACGGTAACGAGAGAGAAGAAGTGGAGATCGAGCC[A/C]AACTCAGTCTCCCTCGACCGAATGATTCTCAGCTACATCG	PREDICTED: hypothetical protein [Vitis vinifera]	4.00E-56
SNP_4697	AATTGAAGCACTTGGCCGGGCAGGAGAGCTTCTTCTTGTA[G/A]CAAGGACCGCGATCCGAGCAGACAACTGTAGGCCGGATGA	hypothetical protein LOC100499725 [Glycine max]	4.00E-10
SNP_5441	CCGGGTGAATGTTCTGATTCAAAAGTGCTTCCATCTTATC[G/A]TATGTACCCTCGACTGCAAGCATTCCCATATAATTGAAAT	PREDICTED: hypothetical protein [Vitis vinifera]	7.00E-43
SNP_12513	AGGTCTCGACAACATCTCCGGCTTCCCAGTCGCTGAATCC[G/A]TCCACTCCGATACCGCACTCAAGTCCAGCACCAACCTCTT	Os05g0575300 [Oryza sativa Japonica Group]	2.00E-29
SNP_6398	CTGTCTTCATGGATGCTAGTAAGCATTTCTCCATCCGCTC[G/A]TCATGACTAATCTGCTGCTCTCGGCACTCCAGGCTGCAAA	predicted protein [Populus trichocarpa]	4.00E-26
SNP_8141	CTTCAAGTACAGAAATTGTAGCCGGTGTGTAATCAAGCTG[C/G]CGAGCACCATGGTTGGAGACAATTATTCCAGATACTCCTG	Aldolase-type TIM barrel family protein [Arabidopsis thaliana]	2.00E-41
SNP_13066	TATCTTTCTATGGCGCAAGTCAGCTCAAGTTCCATTGGGT[A/G]GAGCGTTACGACAAACTATTGGTTGGCTCGGTCCTGTGCC	Os01g0277700 [Oryza sativa Japonica Group]	2.00E-73
SNP_4139	TTCCAGGCCTTGGAGTATTGCACTTCCTCATGTTACAGAC[G/A]TTTCGAAATGAGAAGTTGTTGTTGCCACAATTCGGACATT	predicted protein [Hordeum vulgare subsp. vulgare]	2.00E-41
SNP_8020	CTTCCAAAACCGTCACGCTGTTCGGGATGTTCACCACCTT[G/A]TTATAGCGAGAGATCTTTGTTATAAAGTTGCGCGGATTAT	UDP-glucose 4,6-dehydratase [Arabidopsis thaliana]	2.00E-32
SNP_13041	TCCGCCGAGCTTCACGCCGGCCTGTTAGAAGGACCATCTG[G/A]AATCCAAGCCCGAGGAGATCGTTGTACAACTTCAGGCTAG	putative Acid phosphatase precursor 1 [Oryza sativa Japonica Group]	6.00E-43
SNP_12792	AGGATATCATCCGGACTTTGAAAAATCGCGATGCCTCTTC[A/G]TAGTGAAGAAGGATGGAGAGCTAGAGGATATCTCATATTG	PREDICTED: hypothetical protein isoform 1 [Vitis vinifera]	5.00E-44
SNP_14671	GCCTGCCAGAAGCTCGCATGCATTCTATTCCTCTTTCATC[G/A]TGAATTTCAGGTCCACCAGCCATTCCAGCTAACTTTGGAG	methionine aminopeptidase, putative [Ricinus communis]	9.00E-76
SNP_5662	AAGGGTTCCACAAACCCCAAACATCTTTCGATTATCTACT[C/A]AGCACACAATACTGGCTACCCAAAATACTCTCACCTTAAA	no blast hit	
SNP_13237	TGATGCCGTCCTCGATATAGACATCCGCAACCTGTTGATG[G/A]TGAGCGTTCACGACAGTCCCGCCGGTGATCAAGATCTTGG	hypothetical protein OsI_04141 [Oryza sativa Indica Group]	1.00E-18
SNP_5262	GAGACATGGCTTTTATCACCAGTAGGACCATCTTTGGCCA[G/A]TCCAAAGTCGGAAAGCTTCGCATTATAGTCCGAATCAAGG	Protein kinase APK1A, chloroplast precursor, putative [Ricinus communis]	5.00E-63
SNP_5253	AGTCAATACTATGGACTCCAACAACTACACATGTTGCACC[A/G]CTCTATGCTGCGGCGGATGTTTATGTGATAAATGCCCAGG	predicted protein [Populus trichocarpa]	7.00E-55
SNP_9564	CCTCTGAGCCGAGAGATTCGACGATGTCGTCGCTGGGGAC[A/G]TCGAAGGGGATCCTCGAGATCGCCAAGTTCGCCGTCTACG	no blast hit	
SNP_4658	CCCCTAAGCTAGCACTGTCGGGGATGATGGACAAGCAACA[A/G]ACAGTTGTCAATGAAACAGAGACAAGCCACTCTACAGCCC	no blast hit	
SNP_17302	TGGCATGTCTCGCGGTGTTACCATGGGATATGAAGGTTGC[A/G]CGCTCTTGAAAAGAGGTGTAGATGAAACCAAGGTAAGGGT	OSJNBb0089B03.6 [Oryza sativa Japonica Group]	5.00E-92
SNP_17715	CGCCCTGGATCTTGTCCCCCTCGAATGAGAGCATGGAGCC[A/G]TCCTGGTAGAGGTTGACCAGCGCCGCCCGGTTGGAGTCGA	predicted protein [Populus trichocarpa]]	9.00E-26
SNP_4862	TCAATACTTTTGCACTAGTGTAATTCCAGAATATGTCAGG[C/A]ACTAGTACATCTTGTTTGTCTGCATATAACTTCTTAAATC	unnamed protein product [Vitis vinifera]	2.00E-86
SNP_13755	TTGCAGCAACTTTCCTGCTATCATCCTTTGTTGAGGAGTT[G/A]TTATAGAGCTCATACTGATCTTTCCAGACACGAGAACACA	unnamed protein product [Vitis vinifera]	1.00E-51
SNP_13650	TCAGACTATCATCCCTGCTGGTTCCTCGACATTTGTCTCA[A/G]AGAAGAATCAGGTTACCAGCAAGGAATTATCATCCCAAAC	PREDICTED: hypothetical protein isoform 1 [Vitis vinifera]	6.00E-42
SNP_4388	CCTTTCCTTTCTTGTTAGTGAACGTCCGGCGGTACGTGTT[G/A]ACATACTTGCGAACATCATCGTCTTTGTTAAGGTCGAAGA	S6 ribosomal protein [Asparagus officinalis]	2.00E-100
SNP_9770	TCTCGCAGGACTCGATATCATGCAGAGTTACTACCCTGAA[A/C]GGCTCGGAAAAGTACTTTTAATCCACGTGCCTTATATGTT	Os03g0850700 [Oryza sativa] >gb|AAO20076.1| putative phosphatidylinositol /phosphatidylcholine transfer protein [Oryza sativa Japonica Group]	4.00E-42
SNP_9129-1	AGGATCTGCTGAAGAATGTGTACAATATGCCTCCGAAATC[A/G]GAAGGACAGCAGTCGAACGATCATCGACCCCTTAGCTAGC	unnamed protein product [Vitis vinifera]	5.00E-19
SNP_8033	ATGAGACACGGGGGAAGTTTGAAACAAGATTATATTTCTC[C/G]GCGGTCAAGCAATCCAAAGAGTCAACATAGTCATAAATTG	unnamed protein product [Vitis vinifera]	1.00E-24
SNP_17784	CTGCTGGCCATGTTGCAATCGAATGACATCCATTATCTAC[A/C]ACCAGAACAGGGGTGGCTGCTGTAGGGAATCCTTGTCCGA	no blast hit	

SNP_16951	TTTAGTGGAAGCATCCACCGAGGAGGACTTCACAGGAAAA[G/A]CTGGACAGTCAACAATTCTCCGGCTTCCAGGTTTGGGTTT	Os02g0794700 [Oryza sativa Japonica Group]	6.00E-77
SNP_4988	TGTCTGGGTCAAAATTTTGCAAAAATGTTTGAGATAGCAT[A/T]TGAGAATGAGAAGGGTGAAAAGGCTATGGTTTGGCAGAAT	aminoacyl-tRNA synthase [Oryza sativa Japonica Group]	9.00E-144
SNP_17165	TGGGGATGAGCCTGAAGACTACTGAGTTCTGGTCTTGGGC[G/A]TTCATAATTGCTTTTATGATATTCTGTATGTTGGATTATT	tubulin alpha-1 chain [Oryza sativa Indica Group]	1.00E-113
SNP_12571	AGACATGGACAAAAGAACATTTGAAGGAGAAGTTGAAGTC[A/G]TATGGCGTCGAAAATATTGAGGATTTGACACTAGTTGAAG	putative heterogeneous nuclear ribonucleoprotein [Vitis vinifera]	8.00E-61
SNP_17287	TGATCAAGTGGGCCCCAGAATAGATCATCATCAATTTTAG[A/C]TTTGGAATGAGGCACAGTGAATGAACTTGTCTGCATAGGG	conserved hypothetical protein [Ricinus communis]	1.00E-32
SNP_7534	CCTTGAAACGGATCGTTTGGGATGGGGTGCATTACACAGA[A/G]GCTGCGAACAATTGGGTGTTCGATCAGATCGCTGAAGGCA	early nodulin 8 precursor-like protein [Oryza sativa Japonica Group]	7.00E-22
SNP_13441	GGTCGGGGAAGTCGACGTTGCTCGACTCTCTGGCTGGCAG[A/G]CTGGCCACCAATGCGTTCATGTCTGGGTCAGTGCTGCTCA	white-brown-complex ABC transporter family [Populus trichocarpa]	9.00E-51
SNP_6899	GCAACCATGTCTTCCACCTGCGGTAACTGCGACTGCGCCG[A/C]CAAGAACCAGTGCGTGAAGAAGAGCAACAGCTTCGGCGTT	metallothionein type I [Fritillaria agrestis] >gb|ACC38380.1| metallothionein-like protein [Lilium formosanum]	5.00E-14
SNP_10602	CAGATCGATCCGGTGCGGACGTAGATGGTGTAGACGCATT[C/G]GTTGTCATCATTGGCCTTCTGGAGAGAGATGGCGGTGGAG	dehydration stress-induced protein [Brassica napus]	2.00E-07
SNP_3673	TGGTGAAAGCTATGATGTTATCAGTTCCAGAGAGCTGGGC[G/A]AAAGGGTGGTCCTCCTTGTATCTCCTCAACTCCACCACTC	aspartokinase-homoserine dehydrogenase [Glycine max]	4.00E-33
SNP_6612	GCTGCACGTTTTGAAAGATCTGCTTGTCTTCCATTCCAAG[G/A]TTCCCTACTGCAACACCCATGCAAAGAACTTTCTTCAGCT	Os08g0558800 [Oryza sativa Japonica Group]	5.00E-36
SNP_13728	CAAACTTAATCGAAACATCAGCGAGAAGTGGAGCTGGAGT[A/G]CTCGACCTCCAGGGGAAGTTAACAGACGACATTGAGTGGC	hypothetical protein SORBIDRAFT_01g028430 [Sorghum bicolor]	6.00E-15
SNP_9761	CGGCCATCAAGGCTGCAATTGTTAATTCTCAAACACTTGA[A/G]GAGGTCGCCAGGCTTGAGAAGGCTTTGCAGTCTGGCCAAC	U2 small nuclear ribonucleoprotein A, putative [Ricinus communis]	1.00E-23
SNP_12928	CGTCTTGTGCTCGAAATCCCCAAAAGTTGAGTTGAAGAAA[C/G]TTGCATGTTTAGTTGCAGTCCCTACCGGAACTACCAAAGC	calcium dependent protein kinase 16 [Populus trichocarpa] >]	9.00E-17
SNP_8554	TCATATCGATTCCCGCGAGGCCTCGGGGGTATCTCCAGAG[G/A]TACCTCCTCTTTCTTTTCGAGCTTCTTATCATCACTCAAC	predicted protein [Hordeum vulgare subsp. vulgare]	5.00E-21
SNP_5607	CTTCCCTAACCACAACACTCCCGGAATTATTCAAGTCGAT[G/A]CCTCCTAATGTAAGATTCACTTCACCACCTCTCTGGGGTA	pleckstrin homology (PH) domain-containing protein-like [Oryza sativa Japonica Group]]	4.00E-98
SNP_5094	TTCCTTCTCCTGAGACTGAAGAGATCCATTATGAGAAGGG[A/G]CGCTTTCCGACCATCAACTCCGATGAGATTAATGAGAATC	CCR [Lilium hybrid cultivar]	9.00E-42
SNP_14077	CTGCGGTAAAAGCCTTCTTGCCTCCACCGGTTGATCCAGC[G/A]GGAACAACCTTCAACACATTAAAACGAACAGTCTTTGCCA	40S ribosomal protein S11, putative [Ricinus communis]	5.00E-67
SNP_8370	TCAACTAACATGGGTGGCTATGAACATCCTCTACTCGTAT[C/G]CCTAGCAGGGCTTCTGGTTATAGGGAGCCTAGCAACACAC	zinc finger protein, putative [Ricinus communis]	4.00E-15
SNP_15074	GTGAATCATGGCTTTCGCGAACAAGATCGGGAGTCTCAGG[A/G]GCCTATTGAAGCGGTCGGTGACATCGAATCCGTCGCTGCT	hypothetical protein OsI_33054 [Oryza sativa Indica Group]	4.00E-19
SNP_4593	GGTCCTGTACAGTTCAATATCAAAAGGCTTCCATTACAAG[A/G]TGAAGCTGCATTCTGTGGTAAGGCAGCATATTTCTAATAA	hypothetical protein LOC100526948 [Glycine max]	1.00E-47
SNP_17108	TACAAGATGTAATTGTCCCAAATTCAGAAAATAATTCCTT[A/G]AGTTCTTCGTCACCTATGTTGTCATCCAAGTTCTTCAGGT	unnamed protein product [Vitis vinifera]	5.00E-38
SNP_13377	AAGTCATCGCCCGCGTCCTCGACGTTGTCAAGAGCTTCCC[A/C]AAGGTCGATCCCTCCAAGGTGACGCCGGACGTCCATTTCC	acyl carrier protein, putative [Ricinus communis]	8.00E-35
SNP_12982	GGTAGATGGTGGCTCCCATTGCTATCTCATTGTTGATCTG[G/A]ATGAAGCCAGAGCACAGTAGGTTGTAGCATCCTGTAGCTT	predicted protein [Hordeum vulgare subsp. vulgare]	5.00E-54
SNP_3752	GGTACAAGTCCAAGCATGATTTGAATGGAAAAGATGGTTT[A/G]TTGTCTTTGATTCGGAAGTTCCCTGAGGGCCTTGCTGTTG	hypothetical protein OsI_33458 [Oryza sativa Group]	6.00E-37
SNP_7239	AATGAAGATTGTAGTTGACTGGTATGAAAATGTAATCCTT[C/G]TCAAATAGATTCACTTTACGGGTCCATTTACGGACACGTA	Probable ubiquitin-like-specific protease	1.00E-24
SNP_4154	AATACTTGCAACATGAGCTCGCTTGAGAATCTCCAAGTTC[G/A]AATCACAATCAAGAAGCACAGAGAGAATGCGCTTATAGTG	conserved hypothetical protein [Ricinus communis]	5.00E-12
SNP_685	TGTAGTTTGTCTTACAAAATTTATTCCATCTCTGTAAGAC[G/A]AGGAGAAATCCCTATTTTATACTTATTTTCGAAAGTTAAA	no blast hit	
SNP_6478	AGCCACCTATCTCCACCTCCTTGCCGAGGTGGTCTTTGAC[G/A]AAGACTGCGACCTGCGCAAGATCCAGATTCAACGCAACCA	ATP-citrate synthase, putative [Ricinus communis]	1.00E-17
SNP_6798	GCCCAGCCGCACCAGCTCGTCGCCGATGCTGAATCCCATG[G/A]CCAGATCAACAAAATTGAGGAATCTAGAAAGAGGGTGAGA	Na+/H+ antiporter NHX6 [Zea mays]	2.00E-29
SNP_8046	TCTTGCTCATCTTCGCGTGCTGGTAGTACAGAACGATCAG[G/A]TACCGGTTGGCGACGTTGAAGCCGGAGAGATGATCGACGG	RNA binding protein [Elaeis guineensis]	8.00E-07
SNP_17252	ATACCTAAAGTTGGCTTGAACTTAGTTCAAGTACCTAATG[A/T]CGTTTTGTTAATCATCAAATCTCACCCTGGAAGTCATGTT	no blast hit	
SNP_13102	AAAACATATCGCATTCCACTAAACACTTGCCAACTTAAGA[C/G]CATCTGCCTTCTCTTGCAGGCTAACAAGCAAGCTGTCAAA	transaldolase [Hyacinthus orientalis]	2.00E-44
SNP_17829	TCCAAGCGAAGAAAACGAAGCTGCAGAAGCAAATTGCAGC[G/A]AAGCTGCAGGTGATCGAAGATCTTCAAGGTCAGATTGATG	no blast hit	
SNP_10369	CTTTCTCGTAACTTGAACCTTCTTCGCTACTTAGTCTTTC[G/A]CTACTGGAAAAATTCAGTTTCTTCCTTGAAAAACTTCCTT	hypothetical protein OsI_19654 [Oryza sativa Group]	6.00E-28
SNP_14350	CGTAGATGCCACTGACCTTGTTGGAGATGGCTTCGTGGGC[G/A]AGCATGTAGTCCTTGTAGAGGAGGTAGAAGCCGGCGGCCG	no blast hit	
SNP_13050	CTTGAGCTGCCAGCGAATGAAGCTCTTCTTTCTTCTCAGC[G/A]GACATTTTCTTGTGCGCCACTCGCAGAGTGGGATAAATAA	unnamed protein product [Vitis vinifera]	1.00E-19
SNP_4267	GAATTTCAGCTAGAACACCCTTAGAATAAGCCTTTGAATT[A/T]TCTCTGAGAATGTGTTTCGCCACAGCAGGATCCGAAACAA	cytochrome P450 CYP97A16 [Zea mays]	1.00E-48
SNP_7201	TCGGCCCCTCTAACTCGACAGATCGAAAGACAGAGGTTGG[A/C]CGGAAATGATTGTGCCGGCACTAGGGATGGCGGCGGCTGC	no blast hit	
SNP_13617	GAGCTTTGACTGCAGTCAGAGTATCAATAGGTTGGAGCCC[G/A]CTGTTCTCTGCCAGAGCCATAGGAATGGAATCTAAGGCTT	PREDICTED: hypothetical protein [Vitis vinifera]	1.00E-34
SNP_17072	CAGCCAAAGGTCCAGCAGCAGCCAGCTTTAGAAGGTCCTC[A/G]CGGTTCGCCACTATACTACTGATTCTTGTAATAGCACCAA	metalloendopeptidase [Zea mays]	6.00E-72
SNP_3509	GTTTCGGAGGATCCTTACGCATTAGGTACAAGAAATCTTC[G/A]GTCAAGATCTTTCCCCTCTTGGATGAGATATCTTGGGCTT	TBP-associated factor 13 [Solanum melongena]	5.00E-42
SNP_6421	GTTGAACAGTTACATAATATGTCAGCTTCCACATTGATTC[G/A]GAAGACTTCAAGACCTTTGACTGCATAGCCTCGTCAATCA	hypothetical protein OsI_10853 [Oryza sativa Group]	4.00E-20
SNP_6009	AGGAGTGGGATTGGGACAATTTTCTGATTTTGCATGCTTT[A/G]ATTGGAGCAATGACATGTTACTTCTTCCCATCAATCAAGG	predicted protein [Hordeum vulgare subsp. vulgare]	5.00E-54
SNP_17357	TTGTAGCCTTCCCCTGCTCCATCAAAATGTAATATGAAAT[A/T]ACCACCTCTTTGACTTCCTGTTGAATGACATCATCACACA	unknown [Zea mays]	2.00E-46

SNP_17089	CCAACGTGAATGATATATTTTATGAAATAGCGAGAAGGTT[G/A]CCTCGTGCTCAGCCAGCTCAGAATCCAACAGGGATGGTTC	GTP binding protein [Cichorium intybus x Cichorium endivia]	1.00E-72
SNP_6701	CCATTCTGCCGACATGGAGAACCGGGAGGATGCATTGGTG[G/A]CTCCCGGGAGTGCGAGCGAGCTCGTGAATTGCGATTTGGA	chlorophyll a/b-binding protein type I [Malus x domestica]	3.00E-22
SNP_6522	CTTCTTTCTGCAATAAGACTGTTAAATTGCCCGCACTTCC[A/G]TCTGCAACATCTGCACAAACTGGCAATAGTGGAAACTTAA	hypothetical protein OsJ_22776 [Oryza sativa Group]	2.00E-17
SNP_4365	AAGATGAATTGAAGAATTTCTTCACCAAGTATGGAAAGGT[A/C]GTGGACCAAGAGATTTTACGTGATCATACCACCAAGCGGT	RNA recognition motif-containing protein [Arabidopsis thaliana]	5.00E-39
SNP_10241	AAAAGCACGTGGGCGTCCCCGAAGGGATGGACGTTGATCA[A/G]CAGTGTTGTATAAGAAATATGCGGCTTAGCTTGAGGTGTA	no blast hit	
SNP_5569	CTGATGGTATTGGGAGTGGTCCTACAGCAGCCTCCTACAA[C/G]AGAAGCTCCAGCCTCACACCATTTGCTAACATAAGAGACA	homocysteine s-methyltransferase 3 [Oryza sativa Group]	4.00E-35
SNP_12754	ATGTTCTTAACTGATGCATAACATTCACGAACAGCACTAG[C/A]CATATAATTATCGTCATCAAGTCTCTTTTTCAGATCGTTG	callose synthase [Arabidopsis thaliana]	3.00E-104
SNP_5149	CAACCAGAAGCTTGTTGACGTTCTCACTTGCATATCGGTC[G/A]ATCTCATTCAGCCATTGCTTCACATTGTTGAAACTCTCCT	protein with unknown function [Ricinus communis]	9.00E-61
SNP_9436	ACATGCAGCAACAGACCGGTCCATAAAACAGGTACTGCAG[G/A]GGCGGAGGAGCCGGTTTCAGCTCCACCACCGGTTCAGGAG	no blast hit	
SNP_13149	CATTGGTCTTTAAGTCTGGATTATCGCTTTTATCCGTGTC[G/A]GAATTATCCTTCTGGGATTCAATTTCCATGCTTCCAGCCC	receptor-like protein kinase 3-like [Glycine max]	3.00E-33
SNP_13296	CTCAGTCCCTTGTACATTTCAGCCCTTCGGACGCAATGTT[A/G]CTCTGGAATGGATTCCTCTGGGACAAGCGAAGTCCGGGTC	B0518A01.3 [Oryza sativa Indica Group]	7.00E-61
SNP_12427	CATGCTCACGCCAGTCAAGGCCAGCATGGCCGAAGGAGGC[G/A]TGGAGAAATTCCTCGACAGTGTGGGACTCATCAGTGGCGA	GDP-mannose 4,6-dehydratase [Nicotiana tabacum]	7.00E-75
SNP_3197	TCATTTCCTCAAGCATCTCCCATGAATCATTATCATTCGG[G/A]CTGGTGATACAGAACTGTTCTACTTTCTCAAACTGATGCA	seryl-tRNA synthetase, putative [Ricinus communis]	3.00E-104
SNP_3760	ACTCCACCTTCTTCTGGTTCTTTTCAGTGAGGGGCTTTTG[C/G]GGCTCGACAGGGAAATACTTTATTCCCACAAGATCACCAA	serine palmitoyltransferase [Brassica oleracea]	4.00E-07
SNP_4408	ATTATGATAAGCAACATAGCGAGGATTTACATCTTGTACC[A/G]CCTCAGGTGAATCATCCTATATATGGTTGGTAGGAAGACC	PREDICTED: hypothetical protein [Vitis vinifera]	2.00E-36
SNP_9129-2	GATGTACTCCAGTATAAAATTCATACCAGTACATCTCTTC[G/A]GAGCGGCTAGCTAAGGGGTCGATGATCGTTCGACTGCTGT	unnamed protein product [Vitis vinifera]	5.00E-19

Chapter 6

Using *Lilium* L. and *Tulipa* L. High-throughput Sequencing Data for Estimating Genetic Distance and Positive Selection

Arwa Shahin[1,2], Marinus J.M. Smulders[1], Freek T. Bakker[3], Jaap M. van Tuyl[1], Richard G.F. Visser[1], Paul Arens[1]

[1]Wageningen UR Plant Breeding, Wageningen University and Research Centre, P.O. Box 386, 6700 AJ, Wageningen, the Netherlands

[2]Graduate School Experimental Plant Sciences, Wageningen University

[3]Biosystematics Group, Wageningen University, Wageningen, the Netherlands

To be submitted

Abstract

Next Generation Sequencing (NGS) technologies now enable estimating the phylogeny of species by using allelic variations of multiple nuclear genes simultaneously. Additionally, availability of huge amount of sequence data generated by NGS allows testing evolutionary hypotheses such as that human selection (through breeding processes) has been imprinted at genome level that can be measured by positive selection (estimated by omega value). This hypothesis together with the usage of NGS data for genetic divergence estimation was tested in two economically important bulbous ornamentals: *Lilium* and *Tulipa*. Breeding in *Tulipa* is older than in *Lilium* and consequently higher omega value was expected. We used sequence data generated by 454 pyro-sequencing of four *Lilium* cultivars and five *Tulipa* cultivars. Twenty gene sequences (contigs) of each genus and another seven orthologous gene sequences between *Lilium* and *Tulipa* were used in this study. Nucleotide polymorphism rate of *Lilium* was twice as high as that of *Tulipa*, on average one substitution per 26 bp for *Lilium* compared with one substitution per 48 bp for *Tulipa*. Neighbor Networks of the nine cultivars from the two genera were constructed using POFAD (Phylogeny of Organisms from Allelic Data) based on allelic information of individual accessions. Neighbor Networks generated by POFAD were consistent with a RAxML (Randomized Axelerated Maximum Likelihood) tree generated using the same data, and with previous studies even though we used only few genotypes. Positive selection (estimated by omega value) measured for 47 gene contigs, did not show an indication of higher omega values for *Tulipa* compared with *Lilium*, and thus our hypothesis of linking breeding history and positive selection was rejected. The relation between genetic divergence and omega value were studies. Omega value drops when sequences distance between species increase. Apart from that, our data showed that genes linked to resistance were associated with directed selection.

Introduction

The preponderance of molecular data applied in plant phylogenetic comes from two sources: chloroplast DNA (cpDNA) and nuclear DNA (notably rDNA or ITS) (Chase and Reveal, 2009; The Angiosperm Phylogeny Group, 2003, 2009). Chloroplast DNA has the advantage of straightforward genetics: haploid, non-recombinant and highly conserved with respect to gene content and arrangement, notably among closely related species (Olmstead and Palmer, 1992), which simplifies its use in phylogenetic studies (Schaal et al., 1998; Small et al., 2004). However, cpDNA reveals only half of the phylogenetic origin of a plant-lineage since it is uni-parentally inherited, which may lead to unresolved positions especially in cases of hybrids. Additionally, cpDNA substitution rates are generally lower compared with nuclear DNA. Nuclear rDNA is bi-parentally inherited and its Internal Transcribed Spacer (ITS) has a higher evolutionary rate than cpDNA (Small et al., 2004). Ribosomal genes exist in hundreds to thousands of copies in tandem repeats and undergo concerted evolution (Alvarez and Wendel, 2003). If not all rDNA copies are fully homogenized via concerted evolutionary processes, rDNA ITS should not be used for

phylogenic studies (Alvarez and Wendel, 2003; Booy et al., 2000; Lim et al., 2001). Álvarez and Wendell (2003) actually disregard rDNA ITS as a phylogenetic marker and recommend that it should not be used anymore for angiosperm phylogenetic reconstruction.

Lilium L. and *Tulipa* L. were ranked among the top seven most popular flower bulb genera (Benschop et al., 2010). *Lilium* is classified into seven sections based on 13 morphological and two germination characteristics (Comber 1949), and into four hybrid groups: Asiatic (A, *Sinomartagon* section), Oriental (O, *Archelirion* section), *Longiflorum* (L, *Leucolirion* subsection b), and Trumpet hybrid groups (T, *Leucolirion* subsection a). Phylogenetic relationships within *Lilium* were reconstructed using molecular markers (Arzate-Fernandez et al., 2005; Dubouzet and Shinoda, 1999; Muratović et al., 2010; Nishikawa et al., 2001; Nishikawa et al., 1999). Most of the species clustered into clades correlating with their morphological classification of Comber (1949), but a few behaved differently. Species of section *Leucolirion* (subsection a and b) that were supposed to cluster closely according to Comber (1949), grouped separately. Species of *Leucolirion* (subsection b) were closer to section *Sinomartagon*, and species of *Leucolirion* (subsection a) were closer to section *Archelirion* in both studies (Arzate-Fernandez et al., 2005; Nishikawa et al., 1999). *Tulipa*, was divided into 2 subgenera based on 35 morphological characters and on nuclear genome size, and subgenus *Tulipa* into five sections named: *Tulipa*, *Eichleres*, *Tulipanum*, *Kolpakowskianae*, and *Clusianae* (Van Raamsdonk and De Vries, 1992; Van Raamsdonk and De Vries, 1995; Zonneveld, 2009). *Tulipa* and *Eichleres* sections are important since the commercial assortment of tulips belong to these two sections: *Tulipa gesneriana* L. (*Tulipa* section) and *Tulipa fosteriana* Hoog ex B.Fedtsch *(Eichleres* section) (Van Creij et al., 1997). *Tulipa* has multi-locus rDNA clusters, so rDNA markers were not useful for a further refinement of the phylogeny (Booy et al., 2000).

Recently, Next Generation Sequencing (NGS) technologies have revolutionized phylogenetic methodology (*e.g.* (Fitzpatrick et al., 2006)). Multi-locus, low copy nuclear DNA sequences have been used in phylogenetic studies since the late nineties (de la Torre et al., 2006; Griffin et al., 2011; Hughes et al., 2006; Sanderson and McMahon, 2007). Nuclear DNA sequence inheritance is bi-parental and there is a wealth of long and independently-inherited genes (Small et al. 2004). Also, the ability to identify heterozygosity within individuals and hybrids (allelic variation) is considered a distinct advantage. Using two alleles instead of one can give, in principle, better estimations of phylogenetic relationships between closely related taxa (Joly and Bruneau, 2006; Liu et al., 2008), or in case of species hybrids.

Although the availability of NGS sequencing data in plants opens the door to phylogenetic studies using a wide set of loci, a fully standardized approach using a generally accepted set of tools to build species phylogenies from them is still lacking (Sanderson and McMahon, 2007). The commonly used techniques for estimating phylogenetic trees from multiple-loci data are: concatenation or 'super matrix' methods (Nylander et al., 2004), super tree construction (Beninda-Emonds, 2004), and gene tree parsimony (Page, 1998). BEST (Bayesian Estimation of

Species Trees) (Liu et al., 2008) and BEAST* (Bayesian evolutionary analysis by sampling trees) (Drummond and Rambaut, 2007; Heled and Drummond, 2010) were introduced to estimate species trees from gene trees and deal with the multi-allelic nature of genes. These programs enable incorporating several genes separately as well as estimates of effective population size and implement the MCMC algorithm (Markov chain Monte Carlo) in order to find a posterior distribution of species trees. In this way concatenation is no longer a problem and differences in mutation rate between genes can be included in the analyses. However, using consensus sequences (as in BEST and BEAST*) in which SNPs between the alleles of a gene become ambiguous (IUPAC bases), leads to discarding part of the available data.

Here we explored NGS data generated for genetic resource development and SNP marker retrieval in *Lilium* and *Tulipa* in both a phylogenetic and molecular evolutionary context. The aims of this genomic comparative study were: to establish a straightforward and simple methodology to use allelic NGS data for estimating genetic divergence among four *Lilium* and five *Tulipa* cultivars. We used the algorithm proposed by Joly and Bruneau (2006) that incorporates allelic variation in the reconstruction of phylogenetic relationships. The algorithm, implemented in the program POFAD (Joly and Bruneau, 2006), converts a distance matrix of haplotypes into a distance matrix of individuals by taking the average of distances between the haploids, which is then visualized in a Neighbor Network (Bryant and Moulton, 2004; Joly and Bruneau, 2006). By using this algorithm we can combine the distance matrices of different loci without the need to concatenate their alleles. A priory drawback with this approach is that we infer unobserved (*i.e.* average) distances and use them as the basis for tree or network building, and that no support measures for inferred clades are available.

Apart from tree building, the availability of a considerable amount of coding NGS data in cultivars allows studying other molecular evolutionary aspects of their genomes, *i.e.* the occurrence of positive selection in genes, possibly indicating the effects of breeding and domestication. Whereas lily breeding dates back about 200 years (Shimizu, 1987), significant breakthroughs are only 50 years old however, starting with the breeding of Asiatic hybrids (McRae, 1998a). It has only been since the 1970's that lily has become, after tulip, the most important flower bulb and cut flower (Lim and Van Tuyl, 2006). Tulips were introduced from Turkey into Europe starting from the 16th century (Pavord, 1999). They flowered for the first time in the Netherlands in 1594 and around 1630, tulips were extremely popular (*e.g.* Tulip mania, Pavord (1999) and see http://en.wikipedia.org/wiki/Tulip_mania). The introduced tulips have thus been grown and bred for four centuries (Ridgeway, 2004; Van Tuyl and Van Creij, 2006). All in all, breeding in *Tulipa* is therefore much older than in *Lilium,* although the difference is relatively smaller in number of generations, since juvenile phase of *Tulipa* (5-6 years) is much longer than in *Lilium* (2-3 years).

We ask here the question whether the difference in their breeding history is reflected in the genomes of these cultivars. Claims have been made that human-driven breeding in effect

compares with strong natural selection, and hence, this should be measurable at the nucleotide/codon level, as was shown in rice and grape by (Myles et al., 2011; Yu et al., 2011). We expect the *Tulipa* comparisons to yield higher levels of inferred positive selection (as estimated by omega value = dN/dS, the ratio of non-synonymous to synonymous substitution rate) compared with those in *Lilium* due to the difference in breeding history. Ideally, we would like to explore these hypotheses in *Lilium* and *Tulipa*, using comparisons with wild relatives as in (Chen et al., 2010; Myles et al., 2011; Yu et al., 2011). However, as wild relatives of the cultivars used in our study were not available, we could only test positive selection among the cultivars, and between the two genera.

It was hypothesized previously that positive selection occurs after evolutionary events such as gene duplications, gene loss, copy number variations, or maybe breeding and strong selection pressure, with some advantageous non-synonymous mutations reach fixation quickly followed by purifying, *i.e.* negative, selection (Lynch and Conery, 2000; Yang and Dos Reis, 2011). If so, we can expect a negative correlation between positive selection and sequence divergence. This is tested in our study by comparing the genetic distances and omega values within and between the two genera.

Materials and methods

Plant material

Four *Lilium* cultivars (Fig. 1A) representing the four main hybrid groups of the genus *Lilium* were used: cv. 'Star Gazer' (Oriental, *Archelirion* section), breeding line 'Trumpet 061099' (Trumpet, *Leucolirion* subsection a), 'White Fox' (*Longiflorum, Leucolirion* subsection b), and 'Connecticut King' (Asiatic, *Sinomartagon* section). Five *Tulipa* cultivars (Fig. 1B) were used: 'Cantata' and 'Princeps', which belong to *T. fosteriana (Eichleres* section), and 'Bellona', 'Kees Nelis' and 'Ile de France', which belong to *T. gesneriana* (*Tulipa* section). Young leaves (500mg) were collected and kept at -80°C upon RNA isolation.

Methodology

RNA was isolated using the Trizol protocol (Invitrogen. Carlsbad, CA, USA), and purified using the RNeasy MinElute kit (Qiagen, Hilden, Germany). The cDNA synthesis, normalization of the cDNA, and adaptor ligation for GS FLX Titanium sequencing were performed by Vertis Biotechnologie AG (Freising, Germany). For 454 sequencing, the cDNAs in the size range of 500−600 bp were eluted from preparative agarose gels (Chapter 4).

Sequence data of the four *Lilium* cultivars (Lily-All) and the five *Tulipa* cultivars (Tulip-All) were assembled using the CLC assembler (Chapter 4). As a result of assembly step, an Ace file was generated for each assembly (Lily-All and Tulip-All, Chapter 4) that contains contigs (*i.e.* the consensus of all assembled ESTs that belong to one locus) and their quality score. The two Ace files (Lily-All and Tulip-All, Chapter 4) were used as starting point in this analysis. Randomly, contigs with high coverage (>100 reads per contig and at least 4 reads for each

individual cultivar, no InDels) were picked for further analysis. Selected contigs with deep coverage were opened with SeqMan (Lasergene, version 8) and all reads that belong to one genotype were exported to separate SeqMan files for editing and allele specification (*e.g.* Bellona_A, Bellona_B). All individual allele consensus sequences for each gene were aligned in SeqMan and trimmed to the same size for all cultivars. BlastX was used for annotation of contig consensus sequences. The number of polymorphic sites for each contig were calculated using TOPALi 2.5 (Milne et al., 2009).

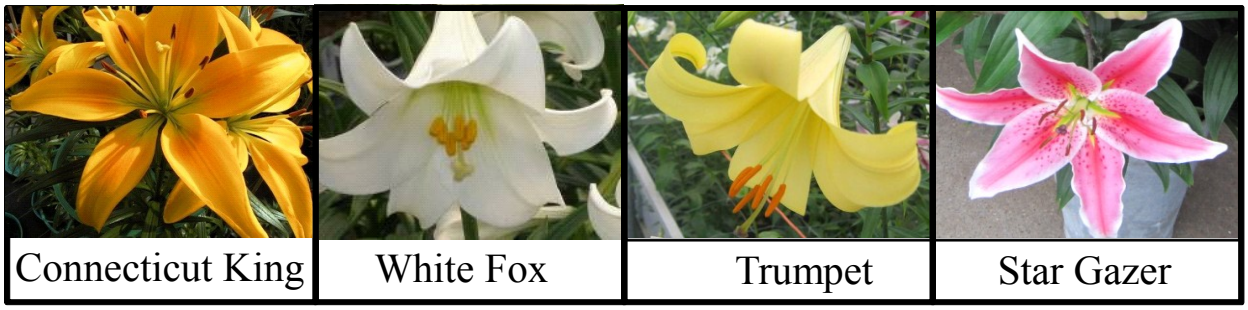


Figure 1A: The four *Lilium* L. cultivars used in this study.

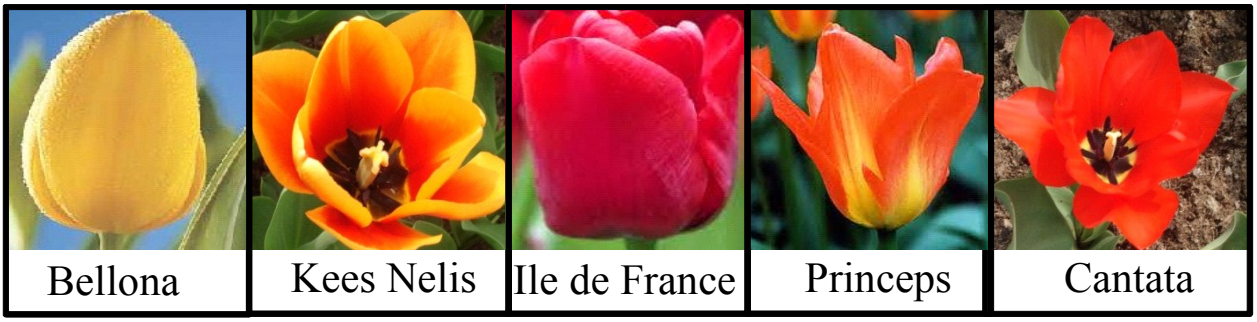


Figure 1B: The five *Tulipa* cultivars used in this study

Additionally, orthologous sequences identified between lily and tulip by blasting contigs of tulip *vs.* lily contigs (BlastN, 1E-20, Chapter 4) were used to choose orthologous genes that have sequences of all the nine genotypes. These sequences were analyzed using the same steps explained above. We expected orthologous sequences between the two genera to be very conservative, and thus we analyzed them separately from the selected gene contigs for each genus, and compared the results for these two sets of results.

Recombination test

To use these gene contigs for tree construction and positive selection detection, a recombination test should be applied to these sequences to avoid using any sequence that is putatively recombined (Vriesendorp and Bakker, 2005). This was done using PDM (Probabilistic Divergence Measure) and DSS (Difference of Sum of Squares) methods implemented in TOPALI 2.5 (Milne et al., 2009). We used default options of the program except for the nucleotide substitution model, where we replaced the (default) Jukes-Cantor model by the

Felestein84 model. Parametric bootstrapping was applied to estimate the significance of the predictions (100 reps). Observed values of DSS and PDM methods beyond the 95 or 99% point of this distribution may well correspond to a recombination event.

Genetic distance analysis

The edited and trimmed alleles of every locus were imported in MEGA 5 software (Tamura et al., 2011). An uncorrected genetic distance matrix (p-distance) was generated for each contig in MEGA 5. Reweighting the individual matrices, which is essential to insure their equal contribution in estimation genetic distance, was done by the algorithm implemented in POFAD (Joly and Bruneau, 2006). POFAD calculates first a pairwise distance matrix of all haplotypes, followed by conversion of this matrix into an organism-level distance matrix, which is then the basis for a Neighbor Network analysis (Bryant and Moulton, 2004). We used both, individual gene matrices and combined matrix to convert allelic variation onto genotype level. The genotypes′ reweighted matrices obtained from POFAD for each gene contig individually, for all the gene contigs of *Lilium,* for all the gene contigs of *Tulipa*, and for the orthologous sequences were transferred to SplitsTree 4 (Huson and Bryant, 2006) to construct Neighbor Networks.

To compare average distance-based POFAD output with a character-based tree-building analysis, we first merged allelic sequences in the seven gene alignments shared between *Lilium* and *Tulipa* by calculating their consensus (including IUPAC bases), and then concatenated the seven alignments, using Mesquite version 2.75 (Maddison and Maddison, 2011). The resulting super-matrix was then analyzed in RAxML (Stamatakis et al., 2008) at the Teragrid of the CIPRES science Gateway (Miller et al., 2010), including the GTR-CAT substitution model (Stamatakis, 2006) and 100 replicates of fast-bootstrapping.

Detection of positive selection

All genes contigs used for this study were checked for possible presence of positive selection using PAML (Yang, 1997). For this, codeml program in PAML package (version 4.4; http://abacus.gene.ucl.ac.uk/software/paml.html) was used to calculate dN (non-synonymous substitution rate) and dS (synonymous substitution rate), and the ratio (ω) between them. Also, transition (Tn) and transversion (Ts) rates were calculated for each contig.

Mesquite version 2.75 (Maddison and Maddison, 2011) was used to ensure that the alignments were in the right reading frame and contained no incomplete or stop codons.

For each contig, branch, site, and branch-site models were tested to estimate ω values (‘model 0′ or ‘1′ combined with ‘NSsites 0′, ‘1′, ‘2′, ‘3′, ‘7′, or ‘8′) using Neighbor Joining trees generated for each contig using PAUP version 4 (Swofford, 2003). The ‘foreground branch′ was ‘Connecticut King′ for *Lilium,* ‘Cantata′ and ‘Princeps′ for *Tulipa*, and *Lilium* for the two genera. The models with the highest log likelihood were branch-site models: ‘model 0′ combined with ‘NSsites 1′, and ‘model 2′ combined with ‘NSsites 3′, with a preference for the first model in majority of the cases. We considered these models as best-fitting our data and use them in our

further analyses (Table 1, 2, 3). The three models with the highest log likelihood values are shown in Table 4 for the 7 orthologous gene contigs.

Furthermore, pairwise omega values, *i.e.* without taking a tree into account, were calculated for all sequences for the seven orthologous sequences shared between the *Lilium* and *Tulipa* data sets, using codeml. This was done to determine if there is a relation between the omega values and sequence distances, and to explore whether *Tulipa* genotypes tend to have higher omega values than *Lilium* or not. Sequence distances were calculated for these contigs by MEGA5 using the Kimura 2-parameter substitution model.

Results

From NGS sequences generated from leaf transcriptomes of lily and tulip cultivars, 52,172 lily gene contigs and 81,791 gene contigs were assembled of which 10,913 gene contigs were orthologous between the two groups (Chapter 4).

For lily, 20 contigs with the highest overall sequence depth were chosen. The length of these contigs ranged between 378 and 957 bp (Table 1). The number of polymorphic sites varied from one polymorphic site in contig_22926 to 71 in contig_36700 (Table 1). There were very few BlastX hits to known genes (Table 1, only the highest hit was shown). Total length of lily sequence data used for this study was 12,485 bp, containing 482 polymorphic sites, *i.e.,* an average of one substitution event every 26 bp.

In tulip, 20 contigs with the highest sequence depth in five cultivars were chosen. The contig length ranged between 219 and 792 bp, and the number of polymorphic sites varied between 4 in contig_74895 and 27 in contig_32233 (Table 2). A total of 10,347 bp with 216 polymorphic sites were available for this study, *i.e.,* a substitution rate of one per 48 bp.

As for orthologous sequences between *Lilium* and *Tulipa*, seven contigs were chosen that contained at least 4 sequences of each of the nine cultivars (Table 3). The contig′s length ranged between 423 and 1,230 bp. The number of polymorphic sites was very low in some contigs (only 8 sites in contig_6081) and much higher in others (200 sites in contig_10364) (Table 3). A total of 5,790 bp with 587 polymorphic sites were available for this part of the study, of which 395 sites were polymorphic only between *Lilium* and *Tulipa*, 124 sites were also polymorphic within *Lilium*, and 68 were also polymorphic within *Tulipa*. This means a substitution rate of one per 47 bp in *Lilium* and one per 85 bps in *Tulipa*, which is almost half the rate of polymorphism calculated above for the 20 contigs in each genus. This could simply be related to the fact that the selected orthologous sequences obtained from both genera are highly conserved genes. Unfortunately, only three of the seven genes had a hit with a known protein in database (Table 3).

Table 1: Description of the 20 *Lilium* contigs used in this study: length, informative sites (calculated using TOPALI), the transition/transversion rate ratio and the omega (dN/dS) values were calculated using PAML.

Contig ID	Length	No. Polymorphic sites	Tn/Ts	dN/dS	Function: BLASTX	E value
Contig_48560	615	11	1.06	0.2	hypothetical protein [Vitis vinifera]	8E-50
Contig_36700	639	71	2.15	4.3	lipoxygenase LOX2 [Populus deltoides]	5E-40
Contig_36290	378	6	17.44	18.8	hypothetical protein SORBIDRAFT_1962s002010 [Sorghum bicolor]	9E-20
Contig_36051	736	12	3.21	0.4	hypothetical protein OsI_19939 [Oryza sativa Indica Group]	2E-125
Contig_35696	551	32	1.14	0.03	unnamed protein product [Vitis vinifera]	3E-43
Contig_34983	660	18	2.35	0.02	hypothetical protein LOC100502367 [Zea mays]	2E-113
Contig_34918	634	41	2.07	0.2	hypothetical protein SORBIDRAFT_02g030210 [Sorghum bicolor]	1E-33
Contig_34429	817	24	2.22	0.1	cellulose synthase BoCesA5 [Bambusa oldhamii]	5E-121
Contig_34202	408	15	2.39	1.04	unknown [Glycine max]	1E-65
Contig_30546	729	49	5.07	Recombinant	hypothetical protein VITISV_004099 [Vitis vinifera]	6E-44
Contig_30305	500	12	3.12	0.4	predicted protein [Populus trichocarpa]	1E-66
Contig_21042	717	39	2.91	0.4	PREDICTED: hypothetical protein [Vitis vinifera]	1E-62
Contig_21012	490	43	1.67	0.8	unnamed protein product [Vitis vinifera]	1E-66
Contig_19882	630	40	2.02	6.33	hypothetical protein OsJ_25146 [Oryza sativa Japonica Group]	7E-61
Contig_19510	372	26	2.92	15.8	unnamed protein product [Vitis vinifera]	7E-15
Contig_6165	714	3	2.01	1.03	ATPase subunit 4 [Citrullus lanatus]	1E-71
Contig_25751	588	11	1.96	0.1	heat shock protein 90-2 [Glycine max]	7E-91
Contig_31438	957	10	2.05	0.4	histone mRNA exonuclease 1 [Zea mays]	2E-94
Contig_22926	510	1	**	2.5	hypothetical protein VITISV_009275 [Vitis vinifera]	6E-14
Contig_20744	840	18	1.48	0.41	Enolase, putative, expressed [Oryza sativa Japonica Group]	2E-133
Overall	12,485	482				

Recombination test

If a gene has a recombination event that means that more than one evolutionary history is present per gene and consequently the weight of non-recombined genes will be down-weighted. Therefore the recombinant gene contigs were discarded from phylogenetic and positive selection analysis. Only one of 20 *Lilium* contigs (contig_30546) showed to have a possible recombination event between positions 157 and 220, thus this gene contig was excluded from further analysis. No recombination event was detected among 20 tulip gene contigs and 7 orthologous gene contigs.

Table 2: Description of the 20 *Tulipa* contigs used in this study: length, informative sites (calculated using TOPALI software), the transition/transversion rate ratio and the omega (dN/dS) values were calculated using PAML software.

Contig ID	Length	No. Polymorphic sites	Tn/Ts	dN/dS	Function: BLASTX	E value
contig_1425	366	20	2.39	0.10	ribosomal protein L32 [Medicago sativa]	2.00E-57
contig_74863	555	7	5.00	0.20	hypothetical protein SORBIDRAFT_03g025100 [Sorghum bicolor]	2.00E-59
Contig_48548	450	8	2.48	5.60	Auxin-responsive protein IAA27, putative [Ricinus communis]	3.00E-30
Contig_54124	468	7	5.51	1.14	unnamed protein product [Vitis vinifera]	0.43
Contig_54419	468	8	13.10	1.08	PREDICTED: hypothetical protein [Vitis vinifera]	4.00E-28
Contig_56016	219	8	7.05	1.50	MYB107 [Arabidopsis lyrata subsp. lyrata]	3.00E-68
Contig_74895	333	4	1.61	14.09	predicted protein [Hordeum vulgare subsp. vulgare]	2.00E-33
Contig_17841	534	10	2.97	0.40	LEM3 (ligand-effect modulator 3) family protein [Arabidopsis lyrata subsp. lyrata]	8.00E-64
Contig_53967	387	11	0.81	1.30	hypothetical protein ARALYDRAFT_484358 [Arabidopsis lyrata subsp. lyrata]	3.00E-58
Contig_48796	465	5	3.84	1.30	fibrillin-like protein [Oncidium Gower Ramsey]	7.00E-71
Contig_17939	453	15	2.08	0.20	Plastohydroquinone:plastocyanin oxidoreductase iron-sulfur	1.00E-110
Contig_48324	558	8	2.63	0.17	hypothetical protein [Vitis vinifera]	7.00E-55
Contig_32233	702	27	1.97	0.33	predicted protein [Hordeum vulgare subsp. vulgare]	8.00E-65
Contig_54223	468	7	6.53	1.90	PREDICTED: hypothetical protein isoform 1 [Vitis vinifera]	1.00E-106
Contig_55287	717	16	4.33	0.07	hypothetical protein POPTRDRAFT_732598 [Populus trichocarpa]	2.00E-127
Contig_54584	711	8	4.43	0.13	unnamed protein product [Vitis vinifera]	2.00E-48
Contig_11296	792	16	5.03	0.04	putative hydroxypyruvate reductase [Oryza sativa]	9.00E-91
Contig_57626	432	19	1.58	7.70	PREDICTED: hypothetical protein [Vitis vinifera]	1.00E-70
Contig_53757	576	7	3.16	0.40	glutamine synthetase [Sorghum bicolor]	2.00E-57
Contig_11279	693	5	1.55	0.06	ubiquitin-conjugating family protein [Jatropha curcas]	2.00E-59
Total	**10347**	**216**				

Genetic distance analysis

Gene trees were constructed for each gene contig separately, and then a Neighbor Network was constructed by combining the weighted genetic distance matrices of all gene contigs using POFAD. The clustering of cultivars was consistent among the 20 gene trees. 'Cantata' and 'Princeps' that belong to *T. fosteriana* grouped together and 'Ile de France', 'Kees Nelis', and 'Bellona' that belong to *T. gesneriana* clustered together as well (Fig. 2B).

In *Lilium*, 'Connecticut King' and 'White Fox' were sisters, as well as 'Star Gazer' and 'Trumpet' (Fig. 2A), in 16 of the 19 gene trees (the exceptions being Contig_25751, contig_6165, and contig_34202; Fig. 3).

Table 3: Description of the seven orthologous contigs between *Lilium* and *Tulipa* used in this study: length, informative sites (calculated using TOPALI software), the transition/transversion rate ratio and omega values (dN/dS) were calculated using PAML software.

Contig ID	length	No. polymorphic sites	Tn/Ts	dN/dS	Function: BLASTX	E value
Contig_6081	987	8	3.2	0.35	unknow protein [Oryza sativa Japonica Group]	3.00E-62
Contig_10364	1230	200	3.8	1.37	transmembrane 9 superfamily protein member 1 [Zea mays]	0
Contig_34202	666	86	2.6	1.27	PREDICTED: hypothetical protein [Vitis vinifera]	6.00E-100
Contig_72799	747	69	1.8	0.07	cellulose synthase [Phyllostachys edulis]	2.00E-115
Contig_36290	423	17	5.6	19.0	hypothetical protein SORBIDRAFT_0070s002020 [Sorghum bicolor]	3.00E-25
Contig_5307	720	87	1.4	1.5	hypothetical protein OsJ_07840 [Oryza sativa Japonica Group]	6.00E-101
Contig_6523	1017	120	2.8	4	peroxisomal acyl-CoA oxidase 1A [Solanum cheesmaniae]	3.00E-166
Total	**5790**	**587**				

The topology that we identified for *Lilium* and *Tulipa* separately was similar to their topology generated using the seven orthologous sequence data of all taxa (Fig. 2C). Using orthologous sequences of both genera was quite useful since we could consider *Tulipa* as an out-group for *Lilium* and vice versa. Notably, the level of polymorphism in *Lilium* showed to be at least two times more than that detected in *Tulipa*, as can be judged from branch lengths (Fig. 2C). The RAxML tree version of the concatenated seven gene alignments, in which alleles had been merged into consensus sequences, showed a comparable topology and branch lengths as found using POFAD (Fig 2D): in *Lilium* 'Star Gazer' and 'Trumpet' (collapsed using POFAD) do form a clade, however without significant support. In *Tulipa*, only the node of ('Cantata' and 'Princeps') 'Kees Nelis' is well-supported (bootstrap value of 94%); those of 'Bellona' and 'Ile de France' remain less certain. Using POFAD the latter two are connected to multiple edges in the Network, possibly indicating 'non tree-like' behavior of the sequences involved.

Detection of positive selection

The ratio of non-synonymous (dN) relative to synonymous (dS) nucleotide changes can indicate whether a gene is under positive or negative selection (Nei, 2005). The criterion for positive/adaptive selection is dN/dS>1, and for negative/purifying selection is dN/dS<1. We used a comparative genomic approach to determine whether the genes used were under positive selection. For that, PAML software was used to compare sequence information of all genotypes included in this study, in various models (branch, site, and branch-site models). Using different models allowed detecting all possible positive selection among different branches and among different codons of the sequences. We are aware, however, that branch models are less powerful compared to site models, *i.e.* the risk of false-negatives is higher under branch models (Anisimova et al., 2001; Yang and Dos Reis, 2011; Yang et al., 2009; Yang et al., 2000).

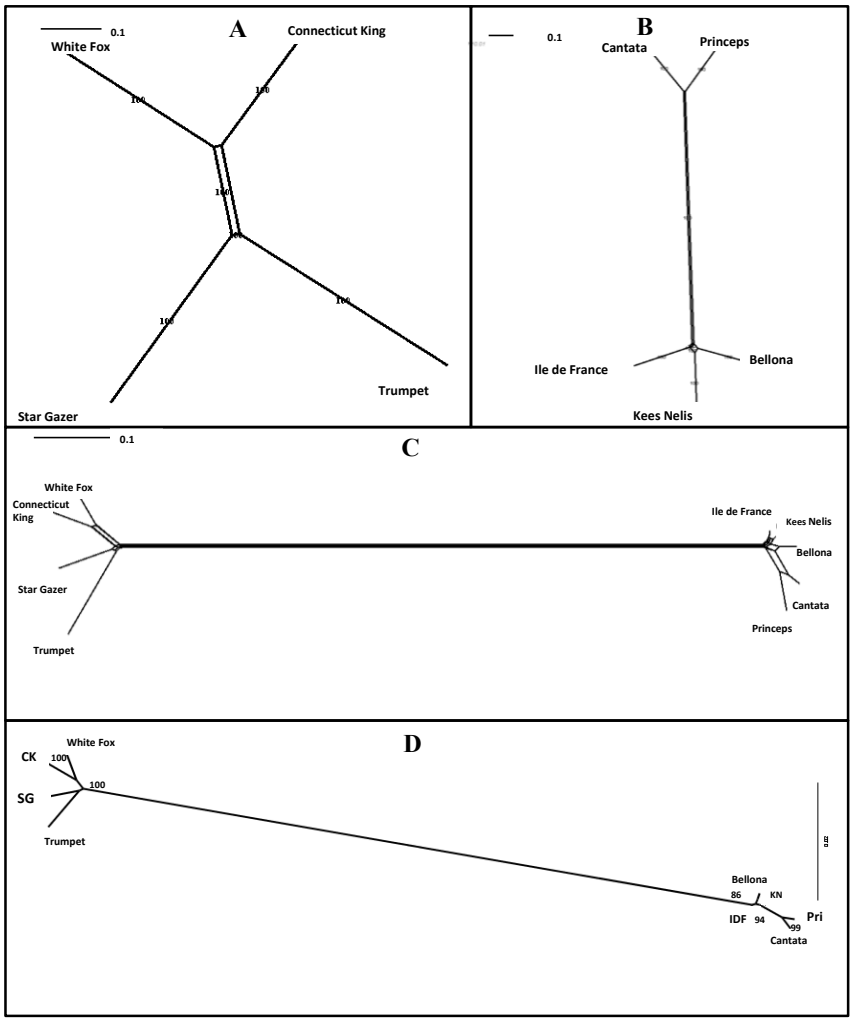


Figure 2: Neighbor Network representation for the relations among cultivars obtained from the combined analysis of all non-recombinant gene contigs. The length of branches refers to the genetic divergence among genotypes. **A**). Neighbor Network based on 19 *Lilium* contigs, **B**) Neighbor Network based on 20 *Tulipa* contigs, **C**) Neighbor Network of seven orthologous contigs between *Lilium* and *Tulipa*, **D**) RAxML tree with scale bar indicating numbers of substitutions per site; bootstrap values (100 reps) indicated at nodes.

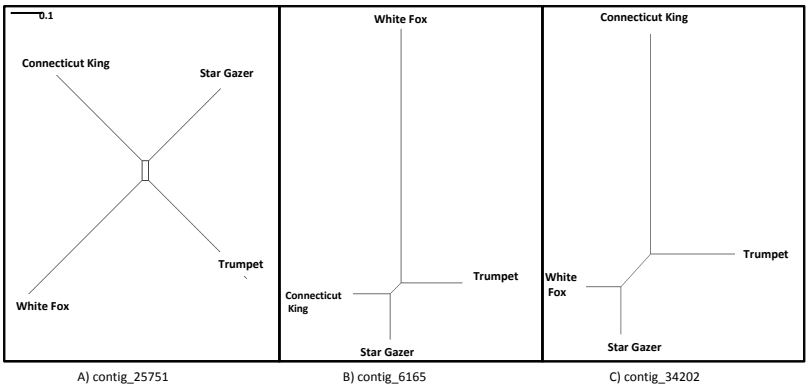


Figure 3: Neighbor Network of three *Lilium* contigs (25751, 6165, and 34202 respectively) constructed using weighted genetic metrics generated by POFAD algorithm.

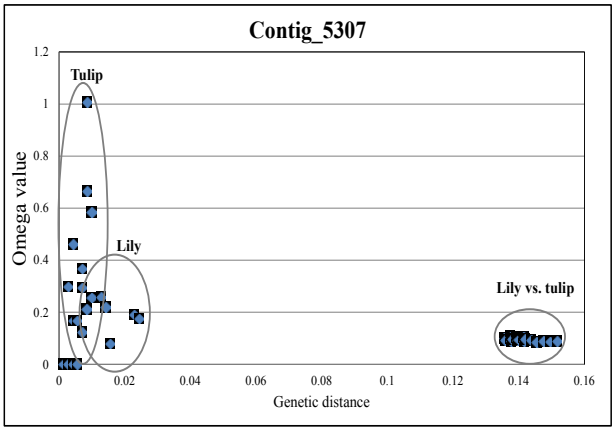
As expected, none of the 46 contigs (19 *Lilium*, 20 *Tulipa*, and 7 orthologous contigs) from the transcriptome showed stop codons. The single recombinant contig was excluded because it may violate the assumption that all sites share the same underling phylogeny (Anisimova et al., 2003). Seven contigs of *Lilium* (Table 1), nine contigs of *Tulipa* (Table 2), and five of the orthologous contigs (Table 3) showed indication of positive selection (dN/dS>1), while the other contigs showed to be under purifying selection (dN/dS<1). Omega value of the gene contigs of both genera appeared to be comparable: ranged between 1.03 and 18.8 for *Lilium* and between 1.08 and 14.09 for *Tulipa* (Table 1 and 2).

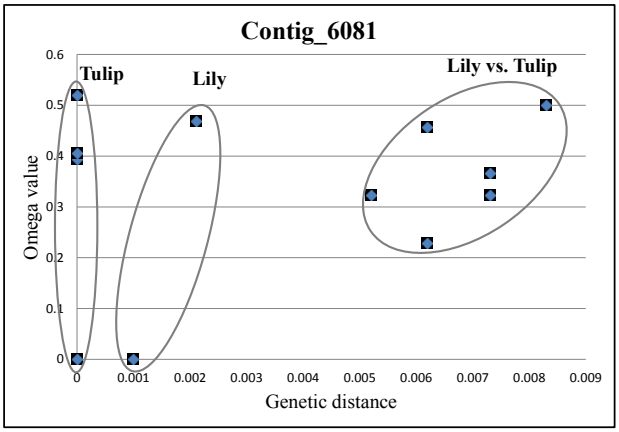
In four out of the seven orthologous contigs, both site and branch site-model indicated omega>1, while branch model did not (Table 4). This might be a consequence of the lower sensitivity of branch models (Anisimova et al., 2001; Yang et al., 2000).

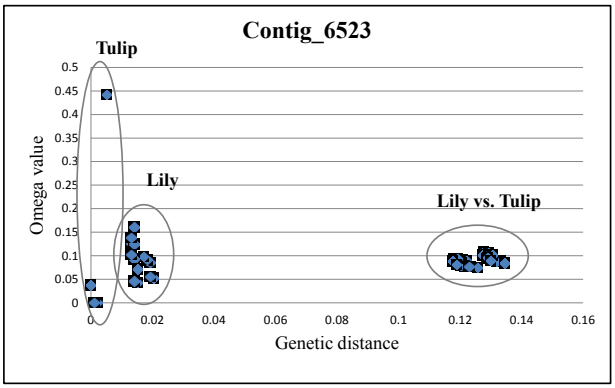
The pairwise comparisons confirmed the presence of positive selection in three contigs (Contig_10364, contig_36290, and Contig_5307), while two contigs, showing omega >1 in the branch site model, were not under positive selection in this pairwise test. Pairwise omega values were plotted versus sequences distances of each gene contig (Fig. 4). We expected to see that the genetic distances within *Tulipa* are lower than those of *Lilium* and that both have lower genetic distances than those between the two genera. While an opposite pattern is expected of the omega values. *Tulipa* is expected to have higher omega value than *Lilium* since it has a longer breeding history (assumed theory in the introduction). The plotting pairwise omega values versus sequence distances for each gene contig showed that (i) gene contigs with omega >1 occur both within and between genera, (ii) the seven genes have largely different omega patterns, and (iii) there is a negative correlation between sequences divergence and omega value (Fig. 4).

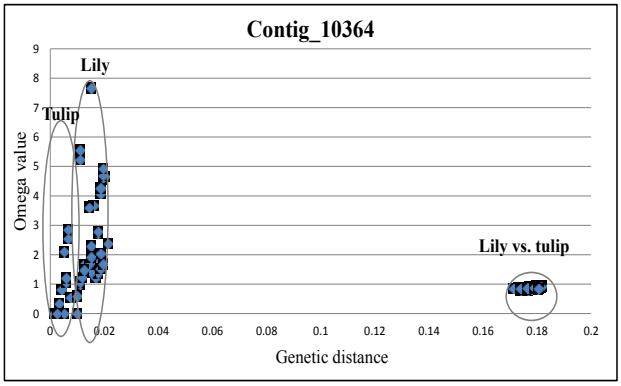
Discussion

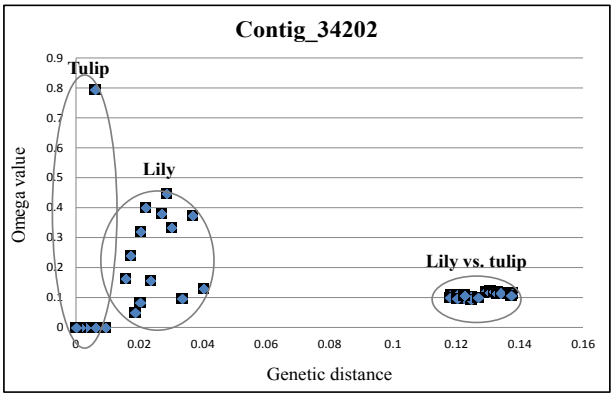
Multi-locus genomic and cDNA sequence data obtained by NGS technology are rapidly becoming one of the main tools for inferring evolutionary relationships (Emerson et al., 2010; Griffin et al., 2011; Scientists, 2009). In this study, we show how to implement allelic information for inferring genetic distances in ornamental plants and thus how they can be used for 'downstream' phylogeny construction. The genetic variation we observed within the genus *Tulipa* was less than that in the genus *Lilium*: one substitution event per 48 bp in tulip compared with one substitution event per 26 bp in *Lilium*. As the cultivars used would reflect overall diversity and classification in these genera (see Introduction) we feel these rate differences are not affected by (taxonomic) sampling artifacts.

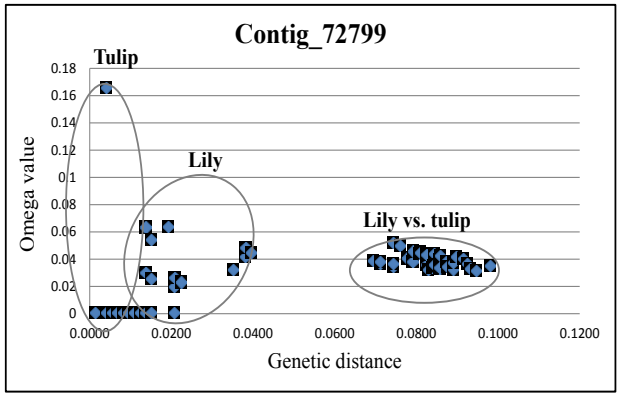












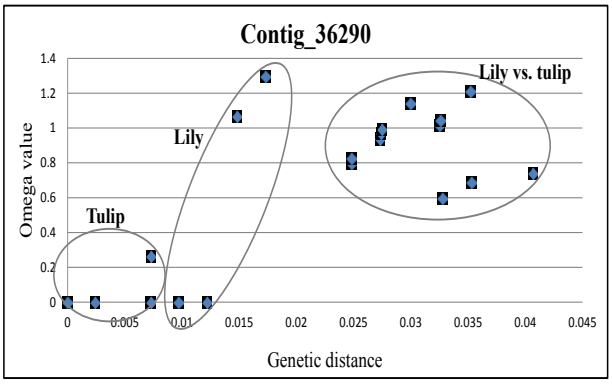


Figure 4: Pairwise omega values (calculated using codeml in the PAML software) versus K2P sequence distance of seven orthologous genes of *Lilium* and *Tulipa* cultivars. Comparisons within *Tulipa*, within *Lilium*, and between the two genera are indicated by circles.

Genetic divergence of *Lilium* and *Tulipa*

Using NGS sequencing, we were able to estimate the genetic distances and build a Neighbor Network for *Tulipa* cultivars using molecular data. The grouping of *Tulipa* cultivars fitted with the traditional classification for this genus, which is based on morphological and cytogenetical characteristics, crossing data, geographical distribution, and genome size (Van Raamsdonk, 1992; Van Raamsdonk and De Vries, 1992; Van Raamsdonk and De Vries, 1995; Van Raamsdonk et al., 1997; Zonneveld, 2009). Cultivars 'Cantata' and 'Princeps' that belong to *T. fosteriana (Eichleres* section) grouped together and 'Ile de France', 'Kees Nelis', and 'Bellona' that belong to *T. gesneriana (Tulipa* section) also grouped together. We could not compare our result with previous studies, since, amazingly, no molecular phylogenetic studies for *Tulipa* were published so far, except for a study about the phylogeny of the order *Liliales* using plastid sequences, in which three species of *Tulipa* that belong to three different sections were used (Fay, 2006).

Lilium cultivars 'Connecticut King' and 'White Fox' that belong to sections *Sinomartagon* and *Leucolirion* (subsection b) respectively were grouped together, while 'Star Gazer' and 'Trumpet' that belong to sections *Archelirion* and *Leucolirion* (subsection a) respectively were clustered together (Fig 2). This is not in agreement with Comber's (1949) classification, based on morphological and germination characteristics; in which 'White Fox' and 'Trumpet' belong to the same section *(Leucolirion*). Similar results were also recorded by other molecular phylogenetic studies (Arzate-Fernandez et al., 2005; Nishikawa et al., 1999). In addition, crosses of *Longiflorum* hybrids (L, *Leucolirion* subsection b) with Trumpet hybrids (T, *Leucolirion* subsection a) are less successful compared with crosses of Trumpet hybrids with Oriental hybrids (O, *Archelirion*) or *Longiflorum* hybrids with Asiatic hybrids (*Sinomartagon)* (Alex van Silfhout, per. obs). Thus, crossability and molecular markers classification appear to support each other in *Lilium*.

Three lily gene contigs presented deviating topologies in our analyses. These reflect either artifacts due to the low number of samples used (long branches and short internode) (Wiens, 2005), or biological deviation which can be explained by assuming that each genomic region underwent an unique array of evolution events such as recombination, mutation or gene flow (Buerkle et al., 2011). If such fragments are highly informative for their own phylogenetic history, it might in principle be possible to track every genomic segment to its origin and thus visualize species hybridization events (Zhang et al., submitted).

The rDNA gene family in *Lilium* genome is present in multiple loci (Lim et al., 2001; Rešetnik et al., 2007) which was reflected in high sequence divergence among the paralogs such as in *Archelirion* and *Leucolirion* sections (Dubouzet and Shinoda, 1999; Nishikawa et al., 1999; Rešetnik et al., 2007). Interestingly, the phylogeny produced using ITS showed in general the same topology as other molecular phylogenetic studies in *Lilium* that used plastid DNA and

isozymes (Arzate-Fernandez et al., 2005; Hayashi and Kawano, 2000) or nDNA (this study). Possibly only one was amplified, or the rDNA in *Lilium* is not too much divergent.

Taxon sampling is considered of utmost importance in phylogenetic studies in order to avoid artifacts such as long-branch attraction and accurately resolve relationships among lineages. Indeed, increased character-sampling combined with incomplete taxon-sampling has been shown to easily result in long-branch attraction artifacts (Wiens, 2005). In this study, we used the minimum number of individuals (one individual per section in *Lilium*) which is, obviously, never sufficient to estimate phylogeny for the genus. Nevertheless, the resulting topology among our cultivars was in agreement with Nishikawa et al. (1999), who used 55 *Lilium* species. This may be related to the availability of a large sequence data set rich with polymorphic sites (more than 12 kb yielded around 500 polymorphic sites in *Lilium*, and more than 10 kb yielded over 200 polymorphic sites in *Tulipa*). Interestingly, similar results were obtained by using our second dataset, almost 6 kb of orthologous sequence of *Lilium* and *Tulipa* species with 587 polymorphic sites. The gene contigs s of *Lilium* and *Tulipa*, which showed to be informative in estimating genetic divergence of these two genera, can be used to study the phylogeny of *Lilium* and *Tulipa* in depth by sequencing those genes in many species per genera and construct gene trees and species trees.

Methodologically, using POFAD helped to include the variation between haplotypes in phylogenetic studies by taking their average (*i.e.* un-observed) distances. However, i) resolution was reduced compared with a consensus sequence approach followed by maximum likelihood tree analysis, and ii) their Neighbor Network does not allow inferring node-support, for instance by bootstrap values. This could be overcome by bootstrapping the sequence alignments, then following the POFAD procedure for each bootstrapped (pseudo)alignment and summarizing the occurrence of groups (bootstrap frequencies), similar to Neighbor Joining bootstrapping. However we are not sure how this POFAD approach could help in the case of hybrids. It might be interesting to run a comparative study between these approaches on hybrids.

Detection of positive selection

Polymorphisms detected from high throughput sequencing a pool of *Eucalyptus* genotypes were shown to be useful in revealing a selection signature among genes (Novaes et al., 2008). Recently, several studies used sequence data either derived from Genbank data or generated by NGS technology, to detect genes that show ability to modify their function as a an adaptive response to natural selection (Barrier et al., 2003; Novaes et al., 2008; Petersen et al., 2007; Roth and Liberles, 2006; Sakai et al., 2011). Availability of sequence data for cultivars in two genera (*Lilium* and *Tulipa*) allowed us to study positive selection in seven genes from these two genera, and whether longer breeding could have been resulted in stronger positive selection in tulip rather than lily.

Seven gene contigs of *Lilium* and nine contigs of *Tulipa* had dN/dS>1. In four out of the seven orthologous contigs shared between the two cultivar groups, omega values of *Tulipa* cultivars were higher while in the rest the omega values of *Lilium* were higher (Fig. 4). Therefore, we do not detect a clear signal of selection due to longer breeding. We feel this cannot be due to sampling artifacts since the cultivars used in this study did reflect the breeding history of the two genera, as most of them have been used widely in breeding programs. The absent of a clear signal may be due to the small set of genes used in this study, which might be insufficient to reflect the whole picture of positive selection in the transcriptomes of the two cultivar groups, or due to the difference in generation time of the species involved. *Lilium* breeding is 2.5 xs faster than the *Tulipa* breeding, which might compensate for the difference in the breeding history between the two genera. Meaning that, the differences in breeding time have not been long enough to be imprinted in their genome.

Pairwise omega values were calculated for each pair of sequences in each gene contig. Inconsistency, however, was recorded when two out of the five orthologous gene contigs that showed to have omega>1 in site or branch-site models, were not under positive selection in pairwise omega test. This is somewhat expected since in pairwise comparisons no trees were used to calculate the omega values, *i.e.* no information from the trees (topology and branches length) could support pairwise omega test as it is the case in branch-site and site model calculations. It was claimed by Nozawa et al (2009) that branch-site model produces false positive results because it produced excessive false positives in their simulation experiment; however, Yang et al. (2009) clarified that with sensible use of statistical methods that were used for detecting positive selection this claims could be rejected.

The plots of pairwise omega values versus sequence distance were informative. There was no general preference for higher omega values for one genus compared to the other. Even though there are a considerable number of nucleotide substitutions between these two genera, they appear not to cause functional changes. This was explained in previous studies (Chen et al., 2010; Lynch and Conery, 2000) by assuming that after gene duplication or strong population level changes (e.g., loss of genetic variation as inbreeding), in which positive selection occurs, purifying selection takes place in which nucleotide substitutions increase due to a lack of selection pressure (Chen et al., 2010; Lynch and Conery, 2000). Therefore we would not expect to see a linear correlation between genetic distances and omega values when plotting them against each other which was confirmed in our study. As far as we know, this comparison between sequences distances and the omega values have not been used and presented before in other studies.

Our data allow us to detect genes that show positive selection (dN/dS>1). A similar study was carried out on 304 ESTs of two *Arabidopsis* species, of which 14 ESTs were identified as candidate genes involved in the adaptive divergence using omega>1 as a parameter (Barrier et al., 2003). In rice, genes of *Oryza glaberrima* showed a larger synonymous-nonsynonymous ratio

than *Oryza sativa* suggesting that *O. glaberrima* has undergone a genome-wide relaxation of purifying selection (Sakai et al., 2011). Our analysis showed that 7 contig in *Lilium*, 9 contigs in *Tulipa*, and 5 out of the 7 orthologous contigs have dN/dS >1. In general, positive selection occurs mainly in the genes linked to defense response/response to biotic stresses (Petersen et al., 2007), while genes related to basic biological processes such as translation and ubiquitin-dependent degradation, and histones and ribosomal proteins are highly conserved (Novaes et al., 2008). In our study, many contigs were from predicted or hypothetical proteins (Table 1, 2, 3), which impedes any comparison between positive selection value and function. This was also observed in other studies (Barrier et al., 2003; Cork and Purugganan, 2005; Roth and Liberles, 2006). In *Lilium*, contig_36700 (dN/dS=4.3) is a lipoxygenase that may play a role in wound response and pest resistance (Hildebrand, 1989), while contig_34429, involved in cellulose synthesis, which is part of the normal cell metabolism, was a conserved protein (dN/dS=0.1) which is consistent with the expectation. In *Tulipa*, all the proteins that showed to have a metabolic function such as contig_17939, contig_11296, contig_53757, and contig_11279, or ribosomal protein (contig_1425, McIntosh and Bonham-Smith, 2001) were under purifying selection (dN/dS<1, Table 2). Other contigs that have a transcriptional factor function such as contig_56016, or is involved in the formation of elastic fibers found in connective tissue (contig_48796) had dN/dS >1 confirming what was reported in previous studies (He et al., 2011; Petersen et al., 2007). Similar results were found in the orthologous contigs. It can be concluded from all three sets that our results support previous studies, and that differences in selection can be detected in a random set of gene sequences.

Conclusions

Our study demonstrates the applicability of sequence data generated by next generation technology for comparative genomic studies. The high number of polymorphism sites identified within and between the two genera: *Lilium* and *Tulipa* showed to be an effective tool for measuring genetic divergence and identifying genes associated with directional selection. Positive selection was detected in both genera and the comparisons of pairwise omega values and the sequences divergence showed that the omega values do not increase with the increase of genetic distances.

Table 4: The different models tested with PAML software to measure positive selection in the common genes between *Lilium* and *Tulipa*. The omega value of the most appreciate model to our data (highest –Ln value) were used for our study.

Contig_ID	Model	NSsite	-Ln	dN/dS
Contig-6523	0	0	-2059.5637	0.09049
	1	0	-2055.0765	0.00010 0.00010 0.03769 0.07503 0.19162 0.15849 0.00010 0.37440 0.00010 0.06365 0.09022 0.00010 0.00010 0.12432 0.57319 0.30650 0.00010 0.00010
	2	3	-2055.3523	background w 0.08242 4.00332 0.08242 4.00332
				foreground w 0.08242 4.00332 0.04652 0.04652
Contig-10364	0	0	-2968.4583	1.16799
	1	0	-2946.7170	999.00000 999.00000 0.44155 1.37594 999.00000 0.00010 0.73572 999.00000 999.00000 999.00000 999.00000 0.84852 999.00000 0.00010 1.64721 999.00000 1.23767
	2	3	-2955.6043	background w 0.83296 5.70375 0.83296 5.70375
				foreground w 0.83296 5.70375 440.32332 440.32332
Contig-5307	0	0	-1485.4258	0.10935
	1	0	-1476.1085	0.14022 0.00010 0.14195 0.79445 0.00010 0.00010 0.09041 0.00010 1.50160 0.28605 0.00010 97.34052 0.00010 999.00000 84.59380 0.00010 999.00000
	2	3	-1484.9498	background w 0.04007 0.25196 0.04007 0.25196
				foreground w 0.04007 0.25196 0.00000 0.00000
Contig-34202	0	0	-1456.7634	0.15044
	1	0	-1447.5588	24.85581 0.00010 0.07802 0.17925 0.15511 0.46488 10.07657 10.80488 0.13244 0.00010 0.00010 0.00010 0.00010 0.00010 0.00010 32.85916 16.65161 0.00010 0.00010
	2	3	-1442.3812	background w 0.04092 1.27663 0.04092 1.27663
				foreground w 0.04092 1.27663 76.22558 76.22558
Contig-6081	0	0	-1389.4134	0.44608
	1	0	-1387.3829	0.00010 0.00010 0.00010 0.35612 0.00010 0.00010 999.00000 0.00010 0.00010 0.00010 0.00010 0.00010 0.00010 0.00010 999.00000
	2	3	-1380.876	background w 0.28255 999.00000 0.28255 999.00000
				foreground w 0.28255 999.00000 0.00000 0.00000
Conitg-72799	0	0	-1563.483	0.03206
	1	0	-1557.118	0.00010 0.00010 0.00010 999.00000 0.00010 0.00010 0.00010 0.07697 0.00010 0.00010 0.00010 0.00010 0.11958 0.05678 0.03298 0.00010 0.00010
	2	3	-1563.058	background w 0.02994 0.02994 0.02994 0.02994
				foreground w 0.02994 0.02994 34.53952 34.53952
Conitg-36290	0	0	-707.7415	1.23805
	1	0	-703.6823	87.46721 62.21842 81.50568 78.03141 103.61902 999.00000 999.00000 999.00000 71.64969 68.19155 0.71500 147.94018 81.13423 999.00000 55.15658
	2	3	-701.5894	background w 0.00000 19.08233 0.00000 19.08233
				foreground w 0.00000 19.08233 0.23810 0.23810

Chapter 7

Towards a Better Understanding of Vase Life of *Lilium* L. Flowers

Arwa Shahin[1,2], Alex van Silfhout[1], Francel Verstappen[3], Harro Bouwmeester[3], Jaap M. van Tuyl[1], Richard G.F. Visser[1], Paul Arens[1]

[1]Wageningen UR Plant Breeding, Wageningen University and Research Centre, P.O. Box 386, 6700 AJ, Wageningen, the Netherlands

[2]Graduate School Experimental Plant Sciences, Wageningen University

[3] Laboratory of Plant Physiology, Wageningen University, P.O. Box 16, 6700 AA, Wageningen the Netherlands

Submitted

Abstract

Flower longevity is an important characteristic for ornamental plants. Longevity starts with anthesis and ends with flower senescence. Regulation of senescence in ethylene-sensitive plants is well studied, whereas regulation of senescence in ethylene-insensitive plants is not well studied. Our aim, therefore, was to define regulator(s) of flower longevity in ethylene-insensitive lily flowers. We studied the effect of exogenous application of sugar on flower longevity and the change in abscisic acid (ABA) levels from anthesis to senescence in seven Asiatic lily genotypes: species *L. bulbiferum*, cv. 'Red Twin', breeding lines 891338-27, 891338-25, 891338-12, 891338-1, and 921442-2. Two treatments were used: "Standard treatment" in which 6 inflorescences of each of the seven genotypes were placed in tap water (1 liter) with 8-Hydroxy Quinolinol Sulfate (HQS), and "Sugar treatment" in which 6 inflorescences of each genotype were placed in tap water (1 liter) with sugar (sucrose, 30 g) and HQS. Longevity of each flower, ratio of dry/fresh weight for each flower at senescence, hormones present in flowers at anthesis, and ABA levels at anthesis and senescence were measured in each treatment. Addition of sugar increased longevity of lily flowers (1 through 3 days) and their ratio of dry/fresh weight (0.01 through 0.03). At anthesis, *Lilium* flowers contained ABA, auxins, and gibberellins, but not cytokinin. ABA levels increased (two to three folds) at senescence compared with anthesis in five genotypes (biological replicates). Exogenous application of sugar delayed the increase in the ABA level. These results indicate a role for ABA in controlling flower longevity of Asiatic lily and a vital role of sugar in delaying senescence.

Introduction

Lily is a perennial bulbous ornamental, belonging to subclass *Monocotyledonae* and family *Liliaceae*. Lily is, according to the statistics of Dutch auctions, the fifth most important cut flower, and the second in flower bulbs (Land- en tuinbouwcijfers 2011: http://www.lei.dlo.nl/publicaties/PDF/2011/2011-029.pdf; http://www.bloembollenkeuringsdienst.nl). Longevity of lily flowers is a very important trait since it has a direct implication on the ornamental vase life and thus on commercial value of these flowers. Ornamental value of lily inflorescence is a function: vase life of individual flower and postharvest expansion and opening of buds (Van der Meulen-Muisers et al., 1999). Improving individual flower longevity will extend the longevity of the whole inflorescence and, therefore, improve the postharvest performance of the inflorescence (Van der Meulen-Muisers et al., 1999). Leaf senescence is also very important for inflorescence longevity since it highly affects the appearance of the whole inflorescence. In this study, however, we focused on flower senescence. Many studies were conducted to improve the vase life response of lily using different compounds separately or combined such as: sucrose, gibberellic acid **'**GA3', 8-Hydroxy Quinolinol Sulfate 'HQS', silver nitrate '$AgNO_3$', silver thiosulphate 'STS', 1-methylcyclopropene '1-MCP', ethylene, vitamin C, citric acid and potassium sulfate **'**K_2SO4' (Ballarin, 2009; Burchi, 2005; Burchi, 2011; Elgar et al., 1999; Nowak and Mynett, 1985; Song, 1996).

Lifespan of a flower is terminated by senescence, *i.e.* wilting or abscission of whole flower or flower parts. Senescence is an active process governed by a well-defined cell death program. Flowers are either ethylene-sensitive and senescence is regulated by ethylene (*e.g.* in carnation, *Petunia* Juss. and orchids), or ethylene-insensitive and senescence is not regulated by this hormone (*e.g.* in *Mirabilis jalapa* L., daylilies, *Alstroemeria* L., *Hibiscus*, *Gladiolus* L., *Tulipa* L., and *Iris* x *hollandica*) (Van Doorn and Woltering, 2008). Developmental events taking place during senescence in senescing tissue involve: physiological and biochemical changes. The physiological changes include loss of water, leakage of ions, and transport of metabolites to different tissues. The biochemical changes include: generation of reactive oxygen species (ROS), increase in membrane fluidity and peroxidation, hydrolysis of proteins, nucleic acids, lipids, and carbohydrates (Tripathi and Tuteja, 2007).

The role of ethylene in flower development has been studied to a great extent in ethylene-sensitive species such as *Petunia*, orchids, and carnation (Nadeau et al., 1993; Tang et al., 1994; Tang and Woodson, 1996; ten Have and Woltering, 1997). Remarkably, little is known about the regulation of ethylene-insensitive senescence. The required signal may or may not require hormones as an intermediate signal (Van Doorn and Woltering, 2008).

In lily flowers, the role of ethylene is unclear. Some studies showed that treatment with the ethylene inhibiter STS (Silver Thiosulphate) enhances vase life of Asiatic hybrids lilies (Nowak and Mynett, 1985; Swart, 1981). Other studies, however, found that senescence of flowers is ethylene-insensitive (Van der Meulen-Muisers, 2000; Van der Meulen-Muisers et al., 2001), or that ethylene has a little effect on the vase life of flowers (Elgar et al., 1999). Pistils of lily flower produce ethylene during the initial development stage, but might not play a primary role in senescence processes (Burchi, 2005; Van Doorn, 2011). Asiatic lilies processed through the Dutch and New Zealand auctions, nevertheless, have to be pre-treated with STS. The lack of clear results makes the benefit of treating cut lilies with STS, however, debatable. Abscisic acid (ABA) is a candidate hormone that might regulate senescence in lily. Abscisic acid showed to have a secondary role during flower senescence in ethylene-sensitive senescence such as *Petunia*, rose, and *Hibiscus* (Borochov et al., 1976; Ferrante et al., 2006; Trivellini et al., 2011b), and might have a major role in ethylene-insensitive senescence in species such as daylily (Panavas et al., 1998).

Several non-hormonal substances are known to be involved in regulating flower senescence such as: calcium, polyamines, and sugars (Tripathi and Tuteja, 2007). Exogenous sugars usually delay visible senescence in flowers (Van Doorn, 2004; Van Doorn and Stead, 1994). The effect of exogenous sugars on senescence was accompanied by a delay in the expression of genes involved in fatty acid and protein remobilization (Eason et al., 2002). In lily, supplying $30gl^{-1}$ of sucrose together with an antimicrobial compound (8-HQC) caused an increase in flower vase life (almost double) (Nowak and Mynett, 1985). Correlation between the onset of senescence and lack of

carbohydrates in lily petals was also recorded (Van der Meulen-Muisers, 2000; Van der Meulen-Muisers et al., 2001).

In this study, the effect of exogenous sugar on the vase life of lily flowers is investigated. The hormones present in lily flowers were identified using LC/MS/MS technology, and changes in the ABA concentrations between anthesis and senescence with and without sugar addition were measured. The relation between ABA concentration and senescence was highlighted. Finally, the influence of exogenous sugar on ABA concentrations in the flower was studied.

Material and Methods

Plant material

Seven lily genotypes belonging to *Sinomartagon* section were used: species *L. bulbiferum* (2n=2x=24), cultivar 'Red Twin' (2n=4x=48) and five Asiatic hybrids; 891338-27, 891338-25, 891338-12, and 891338-1 resulting from crossing 'Connecticut King' with 'Orlito' and 921442-2 resulting from 'Fashion' x 'Montreux' (all 2n=2x=24) (Fig. 1 A, B, C, D, E, F, G). Twelve bulbs (size 12-16 cM) of each genotype were used (Fig. 2A). Bulbs of cv. 'Red Twin', Asiatic hybrids (891338-27, 891338-25, 891338-12, 891338-1), and 921442-2 were grown in the fields of Unifarm, Wageningen the Netherlands, and *L. bulbiferum* was provided by Pennings De Bilt, Breezand the Netherlands. Bulbs were grown in a standard pre-fertilized commercial potting soil under tunnel conditions (little control of temperature and humidity, Fig. 2B). No additional fertilization was used and plants were irrigated daily.

Harvest conditions

Lily inflorescences (stems) were harvested at anthesis of the most mature floral bud by cutting the stems at soil level. Leaves on the basal stem (20 cm) were removed. Six inflorescences of each genotype were placed individually in 1 L glass bottles containing 1 liter of tap water and 150 mg of HQS (TCI Europe N.V., Belgium) and hereafter referred to as 'Standard treatment'. Six inflorescences of each genotype were placed individually in 1 L of tap water included 30 g sugar (sucrose/ table sugar) and 150 mg HQS, and hereafter referred to as 'Sugar treatment'. The inflorescences were placed in a climate room at 17 °C, 60 % humidity and a light intensity of 14 $\mu mol.m^{-2}.s^{-1}$ from a fluorescent lamp (TL-D84 36 W, Philips) during 12 hours per day (Fig. 1H). The inflorescences and the individual flowers were labeled, and the vase life was recorded for each flower.

Flower longevity

Flower longevity was defined as the time between anthesis and wilting (start of deformation, Fig. 3A, B) of the flower. Genotype longevity was defined as the mean of longevity of all flowers of each treatment. Flowers were collected at senescence (except few were collected at anthesis for

hormones extraction, see below), weighed to determine the fresh weight, and dried in the oven (120 °C for 24 h.). Dry weight/fresh weight ratios were calculated.

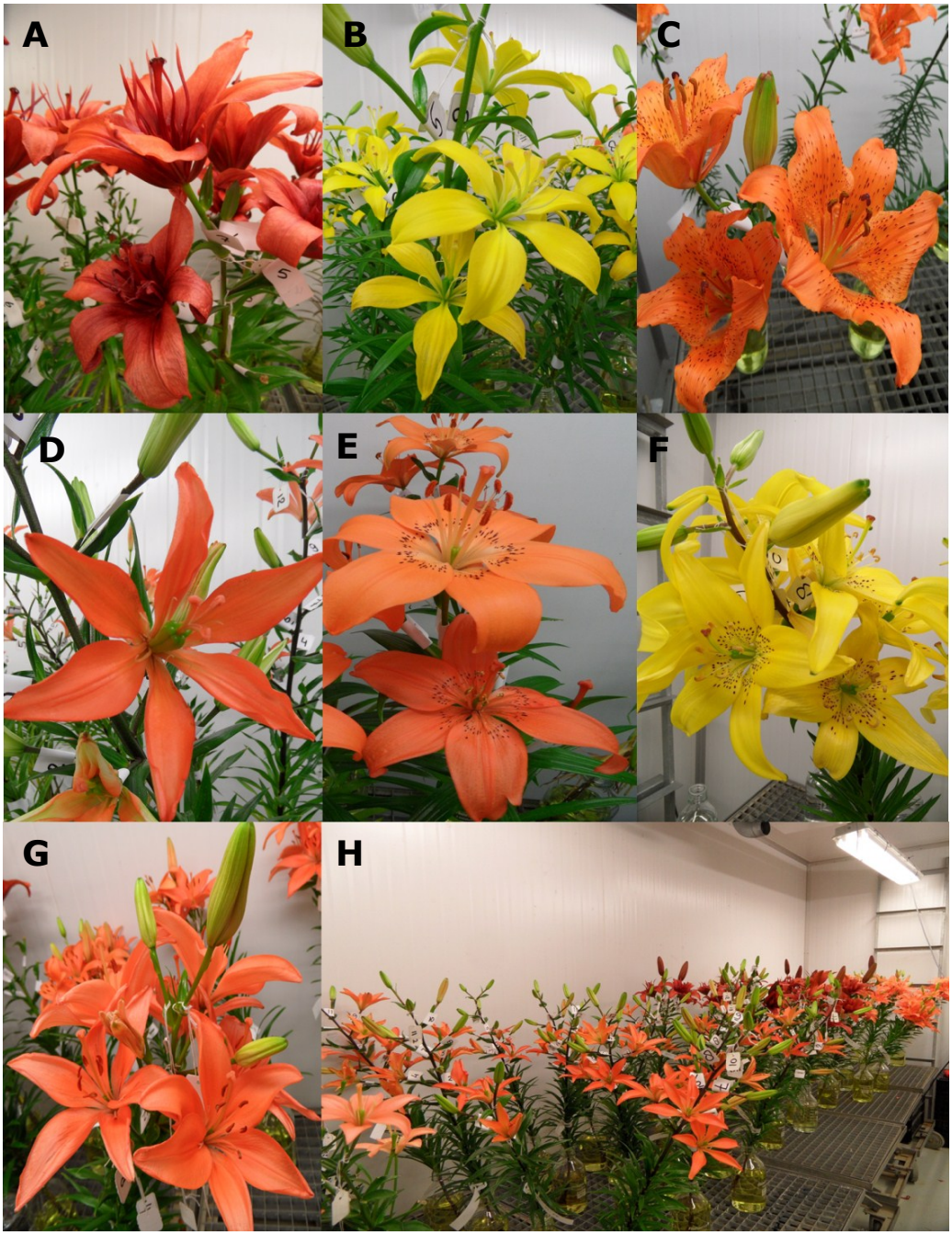


Figure 1: The genotypes used in lily vase life experiment. **A**: cv. 'Red Twin', **B**: 891338-1, **C**: *L. bulbiferum*, **D**: 891338-25, **E**: 891338-12, **F**: 921442-2, **G**: 891338-27 and **H**: the inflorescences of the genotypes in the climate room.

Statistics

Completely randomized designs were used, in which inflorescences (experimental units) were randomized over climate room. Data were analyzed by analysis of variance 'ANOVA' using GenStat 14 (http://www.vsni.co.uk/2011/asides/genstat-14-released).

Figure 2: Genotype 891338-1 bulbs (**A**), and shoots (**B**) planted in a standard pre-fertilized commercial potting soil in tunnel conditions.

Figure 3: Senescence of **A**: cv. 'Red Twin' and **B**: 921442-2 genotypes.

Hormone extractions

To determine hormones presence in lily flowers, levels of ABA at anthesis and senescence, and effects of sugar treatment on ABA concentrations, hormones were extracted. Two flowers of each genotype and of each treatment were collected at anthesis and senescence (Table 1). Collected

flowers were immersed in liquid nitrogen, pooled together and kept at -80 °C untill hormones extraction. Table 1 shows the genotypes, stages and treatments from which two flowers were harvested and pooled together.

Hormones were extracted from 500 mg fresh flower powder (frozen in liquid nitrogen and ground) in 2 mL of methanol containing 2.5 mM diethyl- dithio-carbamic acid (antioxidant) and a mixture of internal standards (IS) to a final concentration of 0.5 nmol/mL. The internal standards were: deuterium labeled abscisic acid (D6-ABA), abscisic acid glucose ester (D5-ABA-GE); ^{13}C-labelled indole-3-acetic acid (^{13}C6-IAA); deuterium labeled cytokinins D6-iP, D6-iPR, D5-tZR, D5-tZ; deuterium labeled gibberellins D2-GA1, D2-GA4, D2-GA7, D2-GA9.

Table 1: The flowers harvested for hormone extractions. Two flowers were collected from the two stages: anthesis and senescence with respect to the two treatments: standard and sugar, at the same time period.

Genotype	Stage	Treatment	days to harvest
L. bulbiferum	Anthesis		1day
L. bulbiferum	Senescence	standard	8days
L. bulbiferum	Senescence	sugar	8days
Red Twin	Anthesis		1 days
Red Twin	Senescence	standard	9 days
Red Twin	Senescence	sugar	9 days
891338-1	Anthesis		1day
891338-1	Senescence	standard	10days
891338-1	Senescence	sugar	10days
891338-25	Anthesis		1day
891338-25	Senescence	standard	9days
891338-25	Senescence	sugar	9days
891338-27	Anthesis		1day
891338-27	Senescence	standard	9days
891338-27	Senescence	sugar	9days
921442-2	Anthesis		1day
921442-2	Senescence	standard	11days
921442-2	Senescence	sugar	11days

Next, samples were sonicated for 15 min in a Branson 3510 ultrasonic bath (Branson Ultrasonics, Danbury, CT, USA) and shacken for 1.5 h. at 4°C in the dark. Samples were centrifuged for 10 min (2500 rpm) and the supernatant was transferred to a 4 mL glass vial. The pellets were re-extracted with another 1 mL methanol and shaken at 4°C for one hour followed by 10 min centrifugation (2500 rpm) and again the supernatant was collected and added to the 4 ml glass vial. The supernatant was vacuum–evaporated and the residue dissolved in 50 µL methanol and 3 ml of water (MilliQ water). The samples were purified using Grace-Pure 100mg C18 columns (Grace Pure C18-Fast 100mg/20 mL SPE, Belgium), preconditioned with methanol and MilliQ water, washed with 1ml MQ-water and dissolved in 1ml of acetone. Then, the acetone was evaporated using vacuum-evaporation and the residue was dissolved in 200 µL of ACN:H2O:FA=25:75:0.1 (ACN: acetonitrile, FA: formic acid). The samples were injected through hand filter SRP4 minisart (Sartorius Stedim biotech, Germany) 0.2 µm filter into LC/MS

vials. Purified samples were loaded to LC/MS/MS machine (Liquid Chromatography-Mass Spectrometry, Waters Acquity UPLC-Xevo, Australia) to measure the hormone concentration.

Hormone detection and quantification by LC- MS- MS

Targeted analysis was performed with a Waters Xevo tandem quadrupole mass spectrometer equipped with an electrospray ionization source and coupled to an Acquity UPLC system 'Waters'. Chromatographic separation was obtained on an Acquity UPLC HSS T3 C18 column (100 × 2.1 mm, 1.8 μm; Waters) by applying a water/acetonitrile gradient to the column, starting from 5% (v/v) acetonitrile in water for 1 min and rising to 50% (v/v) acetonitrile in water in 5.67 min, followed by an increase to 90% (v/v) acetonitrile in water in 1.66 min, which was maintained for 0.67 min before returning to 5% acetonitrile in water using a 0.15 min gradient. In between samples the column was equilibrated for 1.87 min using this solvent composition. Operation temperature and flow rate of the column were 50°C and 0.5 mL min^{-1}, respectively. Injection volume was 40 μL. The mass spectrometer was operated in positive electrospray ionization mode. Cone and desolvation gas flows were set to 50 and 1000 L h^{-1}, respectively. The capillary voltage was set at 3.0 kV, the source temperature at 150°C, and the desolvation temperature at 650°C. The cone voltage was optimized for the Waters IntelliStart MS Console. Argon was used for fragmentation by collision-induced dissociation in the ScanWave collision cell. MRM method was used for identification and quantification by comparing retention times and MRM mass transitions with that of reference standards. MRM transitions for compounds were optimized using the Waters IntelliStart MS Console.

This protocol was applied first to genotype 921442-2 using three samples (anthesis, senescence on standard treatment, and senescence on sugar treatment) in two runs. The first run was to measure cytokinins, and the second run was to measure ABA, IAA, and gibberellins. The hormones that were present in detectable concentrations that also changed during flower development were measured for all other genotypes.

Based on the results of genotype 921442-2, only IAA and ABA were measured. The same hormone extraction protocol was used to extract IAA and ABA for all of the genotypes using only the IS of IAA and ABA. As such the different genotypes form biological replicates. By measuring less hormones per run, results are more reliable and precise.

Results

Variation in vase life of individual flowers

Vase life of seven genotypes in two treatments: standard and sugar were investigated. Notably, genotype 891338-12 showed deviation in its behavior compared with the other genotypes. The last 2-3 flowers to open of its stems fall down without wilting or any other senescence symptoms. This phenomenon resulted in having the vase life of the last 2-3 flowers to open on a stem similar

to or shorter than the first flowers that opened (Fig. 4A). This observation was different from the other 6 genotypes. For the other 6 genotypes, vase life of last flowers to open of a stem increased compared with the first flowers to open (*e.g.* genotype 891338-25, Fig. 4B) for all stems regardless of the treatment. Therefore, genotype (891338-12) was omitted from further analyses.

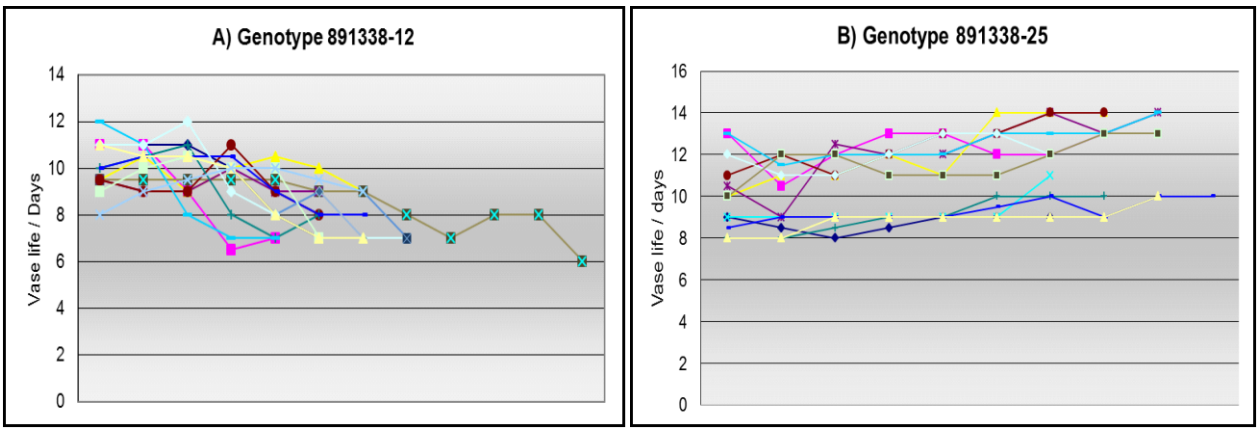


Figure 4: The variation in the vase life of different flowers within the same stem. **A**: genotype 891338-12 in which the vase life of the last 2-3 flowers tends to be shorter than the first ones. **B**: genotype 891338-25 in which the vase life of the last 2-3 lowers on one stem tends to increase. Every line represents a single stem, individual symbols represent single flowers.

The effect of sugar treatment on flower vase life

Vase life of each genotype was calculated as the average of the vase life of all flowers in each treatment (Table 2). In standard treatment, *L. bulbiferum* showed to have the shortest vase life (average of 8.46 days), and 921442-2 had the longest vase life among the six genotypes (average of 12.19 days). Vase life of all genotypes increased with the exogenous application of sucrose (30 gl^{-1}) (Table 2). For instance, vase life of *L. bulbiferum* increased to 10.5 days compared with 8.46 days on standard treatment. ANOVA analysis showed that there was a significant difference in vase life among the genotypes (F-pr <.001), a significant increase in vase life due to sugar treatment (F-pr <.001), and there was no interaction between genotypes and the treatment (F-pr 0.96). Around 53 % of the variance was due to the genotype effect, and 35 % was due to the exogenous application of sugar.

Having considerable numbers of flowers of each genotype for each treatment allowed us to run ANOVA for each genotype separately. All genotypes responded to sugar treatment and their vase life increased significantly, except for cv. 'Red Twin' (Table 2). Vase life of cv. 'Red Twin' increased when applying sugar but the increase was not significant (F-pr 0.338, Table 2). Exogenous sugar explained between 35 to 79 % of the variation in the other five genotypes, whereas in cv. 'Red Twin' only 3 % of the variation between the two treatments was explained by the sugar treatment.

Similarly, the effect of sugar treatment on dry/fresh weight ratios of flowers was tested using ANOVA. There was a significant increase in the dry/fresh weight ratios due to the sugar treatment (F-pr <.001). Also the effect of sugar treatment on the dry/fresh weight ratios of each genotype was assessed, separately. All genotypes, except cv. 'Red Twin', showed a significant increase in dry/fresh weight ratios (Table 3). The explained variance percentages showed that the exogenous sugar treatment explained 15 % to 67 % of the increase in dry/fresh weight ratios (Table 3). This wide variation in the explained variance might be related to the genotypes and their ability to transport, mobilize, and manipulate sugars.

Table 2: The average vase life of each genotype was calculated for the two treatments: standard and sugar (standard error 'SE' is included). The significance between the two treatments, and the explained variance were calculated using the ANOVA. Number of flowers per treatment was included.

	Avg. vase life (standard) ± SE	Avg. vase life (Sugar) ± SE	F-pr	Explained variance
L. bulbiferum	8.46 ± 0.7	10.52 ± 0.9	0.002	56 %
No. flowers	26	21		
Red Twin	9.4 ± 0.6	10.4 ± 1.1	0.338	3 %
No. flowers	57	48		
891338-1	10.36 ± 1	12.3 ± 1.3	0.001	35 %
No. flowers	63	47		
891338-25	9.07 ± 0.54	12.08 ± 1.2	0.001	69 %
No. flowers	54	61		
891338-27	9.19 ± 1	10.7 ± 0.99	0.001	79 %
No. flowers	68	71		
921442-2	12.19 ± 0.7	14.5 ± 0.9	0.001	44 %
No. flowers	66	75		

Table 3: The average ratio between dry and fresh weight for the two treatments on six lily genotypes (SE = standard error). Significance between the two treatments was calculated by ANOVA using the variation between flowers of each treatment

	Avg. dry/fresh weight (standard) ± SE	Avg. dry/fresh weight (Sugar) ± SE	F-pr	Explained variance
L. bulbiferum	0.08 ± 0.02	0.12 ± 0.03	0.007	31 %
No. Flowers	26	21		
Red Twin	0.06 ± 0.01	0.07 ± 0.02	0.127	15 %
No. Flowers	41	35		
891338-1	0.07 ± 0.02	0.14 ± 0.01	<.001	17 %
No. Flowers	59	42		
891338-25	0.07 ± 0.01	0.1 ± 0.01	0.002	30 %
No. Flowers	54	61		
891338-27	0.07 ± 0	0.1 ± 0	0.002	24 %
No. Flowers	58	64		
921442-2	0.07 ± 0.01	0.11 ± 0.01	<.001	67 %
No. Flowers	64	73		

Hormone measurements

Hormone measurements in genotype 921442-2 showed that ABA, auxins, gibberellins, and cytokinins were present in lily flowers at anthesis. Hormone concentrations varied among the three samples: anthesis, senescence of standard treatment, and senescence of sugar treatment.

Cytokinins (N6-Isopentenyladenine (iP), N6-Isopentenyladenosine (iPR), trans-Zeatin riboside (tZR), cis-Zeatin riboside (cZR), N6-Isopentenyladenosine-5-monophosphate (iPMP), trans-Zeatin riboside-5-monophosphate (tZRMP)) were only measurable at very low concentrations at anthesis. Gibberellins (GA1, GA4, GA7, and GA9) were measurable: GA1 was around 0.1 pmol/mg FW (fresh weight) and was slightly reduced to 0.08 pmol/mg FW at senescence in both standard and sugar treatments; GA4 was around 0.02 pmol/mg FW in all three measurements; GA7 was around 0.012 pmol/mg FW in the three measurements; and GA9 concentration was very low in the three measurements (around 0.001 pmol/mg FW). Auxin was measurable at low concentration on anthesis (close to 0.008 pmol/mg FW) and was not measurable at senescence. ABA (both ABA and ABA-GE: abscisic acid glucose ester) was measurable: close to 0.032 pmol/mg FW at anthesis increasing to 0.056 pmol/mg FW at senescence, and to 0.04 pmol/mg FW on sugar treatment. Based on hormone measurement results of genotype '921442-2', only auxin and ABA were measured in the other genotypes (biological replicates).

The effect of sugar treatment on hormones in lily flowers

Auxins and ABA were extracted of all the 6 genotypes (Table 1). Flowers were harvested at the same time point for both treatments. Meaning, if the flowers of *L. bulbiferum* senescenced on standard treatment on the day 8th of their opening, flowers of both treatments were collected at day 8th. This was done even though the flowers will not be wilting before day 10th in the sugar treatment (Table 1). This allowed us to compare hormone levels at the same time point and see the effect of sugar treatment.

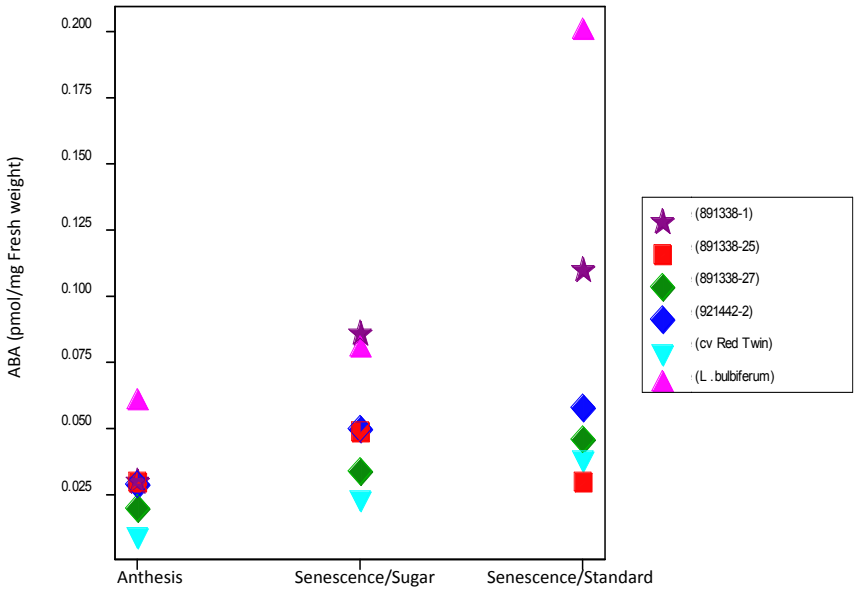


Figure 5: The ABA concentration (pmol/mg fresh weight) for six lily genotypes at two stages: anthesis and senescence (standard and sugar treatments).

The ABA concentrations increased significantly (F-pr 0.020, Table 4, and Fig. 5) from anthesis to senescence in all but one genotype on Standard treatment, which can be considered biological

replicates (Table 4). ABA increased more than three fold in *L. bulbiferum*, 'Red Twin', and 891338-1, and two fold in genotypes: 891338-27 and 921442-2 whereas there was no measurable change for 891338-25 (Table 4). In sugar treatment, ABA concentration increased from anthesis to senescence (Table 4), however the increase was not significant (F-pr 0.264). A strong correlation (0.9) between vase life and ABA concentration was found. Auxins were only measurable at low quantities at anthesis, and the concentration decreased towards senescence where they are immeasurable in most cases (Table 4).

Table 4: The ABA and IAA levels (pmol/mg fresh weight) in lily flowers were measured at anthesis and senescence and under the standard and sugar treatments

Genotype	Stage	Treatment	IAA (pmol/mg FW)	ABA (pmol/mg FW)	Vase life (days)
L .bulbiferum	Anthesis		0.007	0.06	
L .bulbiferum	Senescence	Standard	Not detectable	0.2	8.46
L .bulbiferum	Senescence	Sugar	Not detectable	0.08	10.52
Red Twin	Anthesis		0.004	0.01	
Red Twin	Senescence	Standard	Not detectable	0.039	9.4
Red Twin	Senescence	Sugar	Not detectable	0.024	10.4
891338-1	Anthesis		0.008	0.03	
891338-1	Senescence	Standard	0.001	0.11	10.36
891338-1	Senescence	Sugar	0.0009	0.086	13.3
891338-25	Anthesis		0.005	0.03	
891338-25	Senescence	Standard	Not detectable	0.03	9.07
891338-25	Senescence	Sugar	Not detectable	0.049	12.08
891338-27	Anthesis		0.018	0.02	
891338-27	Senescence	Standard	Not detectable	0.046	9.19
891338-27	Senescence	Sugar	0.006	0.034	10.7
921442-2	Anthesis		0.009	0.029	
921442-2	Senescence	Standard	Not detectable	0.058	12.19
921442-2	Senescence	Sugar	Not detectable	0.05	14.5

Discussion

In this study, the effect of sugar addition on vase life and dry weight of lily flower was examined, the hormones present in the flowers at anthesis were identified, the ABA concentration change between anthesis and senescence was studied, and the effect of sugar treatment on ABA levels were highlighted.

The effect of sugar treatment on flower vase life

Exogenous sugar was found to increase vase life in ethylene-sensitive and ethylene-insensitive senescence (Van Doorn, 2001; Van Doorn, 2004). Sugar addition increased significantly vase life and dry/fresh weight ratios in all studied genotypes except cv. 'Red Twin'. Similar results were found in other studies in lily (Burchi, 2011; Nowak and Mynett, 1985). Comparable in the also ethylene-insensitive tulip, application of exogenous sucrose or trehalose (non-reducing disaccharide consisting of two α-glycosidically linked glucose units) prolonged vase life of tulip

flowers (Azad et al., 2008; Iwaya-Inoue and Takata, 2001; Wada et al., 2005). It is not clear to what extend exogenous sugar improves the vase life of flowers. In ethylene-sensitive senescence, exogenous sugar indirectly decreased ethylene sensitivity as shown in carnation (Mayak and Dilley, 1976). Soluble sugar acts as a repressor of senescence at transcriptional level, and it is more efficient than STS in ethylene signaling inhibition (Hoeberichts et al., 2007). However, the mechanism of action of exogenous sugar supply in ethylene-insensitive flower is not known.

It was suggested that in ethylene-insensitive senescence, exogenous sugar may control petal senescence by increasing the level of osmotic solutes (Van Doorn, 2004). This might be explained by the relation between soluble sugar and osmolality. In lily flowers, complex carbohydrates, mainly starch, are converted into soluble sugars (glucose and fructose) at anthesis (Arrom and Munné-Bosch, 2012; Van der Meulen-Muisers et al., 2001). This means an increase in the osmolality. Osmolality, however, decreased dramatically at senescence in both lily and daylilies (Bieleski, 1993; Van der Meulen-Muisers, 2000), and the total carbohydrate content in the petals of lily flowers decreased at senescence (Van der Meulen-Muisers, 2000). This might be due to the mobilization of the degraded carbohydrate into the phloem. Sugar concentration in the phloem increased by a factor of 14 in daylily at senescence compared to the closed floral bud. Another hypothesis is that sugar starvation could be the direct cause of senescence (Van Doorn, 2004). That was suggested due to the similarity between starvation-induced changes in cell physiology and those observed before cell death during senescence (Van Doorn, 2004). However, at the first symptoms of wilting petals still have satisfactory osmotic solutes. Then again, it is not clear where the sugar is localized in the cell (cytoplasm or vacuole) at the time of petal wilting (Van Doorn, 2004; Van Doorn, 2011). It might be that the sugar is not available to the mitochondria resulting in a decrease in ATP production and cell collapse shortly afterwards.

It was noticed, in our data, that the lifespan of the last flowers to open on an inflorescence was longer than that of the others. If flower buds and opened flowers are present on an inflorescence, an earlier senescence was induced in the older flowers (Van Doorn and Woltering, 2008). On the other hand, competition for carbohydrates among flower buds on the same inflorescence was recorded (Van der Meulen-Muisers, 2000; Van der Meulen-Muisers, 1995). This might mean that, no or less competition for carbohydrate among the last few flowers of the inflorescence occurs and thus their lifespan will be extended.

The ratio of dry/fresh weight of flowers was increased when exogenous sugar was applied. Similar findings were recorded by Van Doorn (2001) where flower size in sugar treatment increased. Sugar delays senescence, and thus gives flowers longer time to grow and increase their sizes and weights.

Exogenous sugar increased vase life and dry/fresh weight ratio in ‘Red Twin’, however, this increase was not significant. ‘Red Twin’ is a tetraploid cultivar and has almost a double size (stem weight and stem diameter) compared with the other five diploid genotypes used for this

study (data not shown). All the other genotypes used for this study and showed significant increases in their vase life are diploid. Meaning that, the amount of applied sugar (30 $gl^{-1)}$ was not enough for this tetraploid cultivar to prolong its vase life significantly. This might indicate that the amount of exogenous sugar that should be applied to extend the vase life of any genotypes might be linked to ploidy level.

Lily flower's hormones

Hormones of lily flowers present at anthesis were measured using a sensitive chromatography method LC/MS/MS. Cytokinins were present in rather low concentrations, which is similar to the recent findings of Arrom and Munné-Bosch (2012). Gibberellins (GA1, GA4, GA7 and GA9) were present in measurable quantities. Auxins could be measured at anthesis although at low concentrations (close to 0.008 pmol/mg of flower fresh weight) but were not measurable at senescence. ABA (both ABA and ABA-GE: abscisic acid glucose ester) were present in measurable quantities. As ABA is expected to play a role in senescence, we focused on measuring ABA levels (Arrom and Munné-Bosch, 2012).

ABA plays a major role in late seed development and adaptation to environmental stresses such as drought and other stress responses (Gazzarrini and McCourt, 2001; León and Sheen, 2003). However, the role of ABA role in flower senescence is not clear yet. Our measurements of ABA concentrations showed 2 to 3 fold increase in ABA concentrations from anthesis to senescence. Similar results were recorded in other ethylene-insensitive senescence species. The ABA levels increased 2 fold (reaching 400 pmol/petal) from flower opening till senescence in daylilies (Panavas et al., 1998). Similarly, ABA concentration increased during petal senescence in cocoa (Aneja et al., 1999). Exogenous application of ABA, hastened flower senescence (Borochov et al., 1976; Panavas et al., 1998) and induced many senescence-related changes such as: lipid peroxidation, protease activity, and expression of novel DNases and RNases in ethylene-insensitive daylilies, cocoa, and *Iris*. Thus, ABA is thought to be the primary hormonal regulator of flower senescence in these flowers (Aneja et al., 1999; Panavas et al., 1998; Zhong and Ciafré, 2011). The application of an inhibitor of the ABA biosynthesis decreased ABA levels and extended vase life of cocoa flowers (Aneja et al., 1999). The results taken all together support the notion that flower senescence of ethylene-insensitive plants is possibly regulated by ABA.

In ethylene-sensitive flowers, ethylene and ABA are in interaction. External application of ABA to flowers induces a large increase in ethylene production and hastens petal wilting in carnation (Mayak and Dilley, 1976). In *Hibiscus* and rose, ABA appeared to be involved in flower senescence but mediated by ethylene (Müller et al., 1999; Trivellini et al., 2011a). The endogenous concentration of ABA in *H. rosa-sinensis* during natural senescence peaked at anthesis stage and decreased during senescence (Trivellini et al., 2011a). Also, ABA was found to have a regulatory effect on ethylene bio-synthesis (Trivellini et al., 2011b). However, the ABA accumulation was considerably decreased in *Hibiscus* flowers at senescence, showing that ABA

may be antagonized by ethylene (Trivellini et al., 2011a; Trivellini et al., 2011b). This interaction is reversed in ethylene-insensitive flowers. At senescence of cocoa flower, ABA levels increased and ethylene levels decreased to undetectable amounts (Aneja et al., 1999; Panavas et al., 1998). In this study, ABA increased significantly with flower development whereas it is known that ethylene concentration at senescence is low (Burchi, 2005). In order to further clarify the role of ABA in senescence and its interaction with ethylene, the function and regulation of ABA induced transcription should be studied. This could be done by developing ABA deficient mutants and by testing specific inhibitors of ABA in ethylene-insensitive species (Zhong and Ciafré, 2011).

Sugar and ABA interaction

Studies on the effects of ABA and sugar on a range of developmental processes have suggested interactions among signaling pathways that may be antagonistic, synergistic or simply additive depending on the processes and on the concentration and chemical form of the sugar signal (Finkelstein and Gibson, 2002; Ramon et al., 2008). In seed dormancy and germination studies, ABA accumulation and transcript levels of ABA biosynthesis genes (ABA1-3, and AOA) are significantly increased by glucose (Cheng et al., 2002; Ramon et al., 2008). In leaf senescence, however, this was not the case (Pourtau et al., 2004). ABA and sugar, applied independently, triggered leaf senescence and ABA is not required for sugar-dependent induction of leaf senescence. In flower senescence the crosstalk between ABA and sugar seems different. In our study, application of sugar delays the increase in ABA levels during flower development compared with the standard treatment. This seems similar to the effect of sugar in delaying protein and lipid degradation (Van Doorn, 2004). Sugar and ABA showed to have opposite effects in rose flowers (ethylene-sensitive). While sugar prolongs vase life of rose, ABA in the presence of ethylene shortens it (Borochov et al., 1976; Müller et al., 1999). A similar relation was found between sugar and ethylene. There is an antagonistic relationship between glucose and ethylene signaling pathways (in *Arabidopsis*) at flower senescence (Zhou et al., 1998). Overall, the availability of sugar decreased ethylene and ABA levels at flower senescence. This might indicate that sugar, ethylene and ABA might have the same signaling pathway, or interacting pathways.

Several studies were conducted to explain the genetic interaction between sugar, ethylene, and ABA using mutants available in *Arabidopsis*, however, only seed development and stress responses mechanisms were highlighted (León and Sheen, 2003; Ramon et al., 2008). The availability of sugar-signaling mutants in *Arabidopsis* have uncovered many unexpected links between sugar and plant hormone signaling (León and Sheen, 2003). Sugar-insensitive mutants (*gin1*, *gin5*, *isi4,* and *sis4*) contain low endogenous ABA levels compared with wild-type plants (Arenas-Huertero et al., 2000; Zhou et al., 1998). The ABA-deficient mutants (*aba1*, *aba2*, and *aba3*) display the sugar-insensitive phenotype (Arenas-Huertero et al., 2000; Zhou et al., 1998). This finding suggests a close link between ABA and sugar (León and Sheen, 2003). On the other hand, ethylene-insensitive mutants (*etr1*, *ein2*, *ein3*, and *ein6*) display glucose over-sensitivity, and partially suppress the biosynthesis of ABA (Cheng et al., 2002; Zhou et al., 1998), which

suggests an antagonistic relation between ethylene and sugar signaling, and ethylene is negatively regulated by the ABA level (León and Sheen, 2003). These results cannot be applied for senescence, since as we explained previously that the relation between ABA and sugar varies due to developmental stage. Also, these studies were carried out using mutants of an ethylene-sensitive plant (*Arabidopsis*), and we are not sure if similar results will be obtained using mutants of an ethylene-insensitive plant.

Conclusions

In our study, we confirmed the important role of sugar in prolonging the vase life of lily flowers. The ABA concentration increased dramatically at senescence compared with anthesis which might reflect a regulatory role for ABA in controlling vase life in lily. Interestingly, sugar treatment decreased ABA concentration, suggesting opposite effects of ABA and sugar on lily senescence and a possible crosstalk between the pathways of both. It is important to understand the different mechanisms that control vase life in lily which will have a direct implication on the commercial value of this flower crop.

Acknowledgments

We would like to thank Tatsiana Charnikhova for her support in the lab, Dr. Yury Tikunov for the nice discussions, and Dr. Chris Maliepaard for the statistical support.

Chapter 8

General Discussion

Even though *Lilium* is an important ornamental monocot perennial bulbous crop, its genomic resources are limited. This thesis aims to establish genomic resources for molecular breeding and other genetic studies in lilies with focus on constructing linkage maps, identifying QTLs and generating DNA sequence resources using next generation sequencing (NGS) technology that support mapping and other genomic studies such as comparative genomics. Also, several of important ornamental traits that are of economic importance have been studied such as fertility, flower color, flower spots, flower direction (up/down-facing), and stem color as well as lily flower longevity as measured by vase life performance.

This chapter will discuss the major aspects dealt with in this thesis. The reasoning of the chosen methodology, the potential alternative approaches in lily breeding, and the future possible applications of the established resources will be discussed.

Molecular breeding

The aim of many plant researches is to explain natural phenotypic variation in terms of simple changes in DNA sequence. Plant breeders aim to use these functional changes in the DNA to improve their breeding programs. Conventionally, linkage mapping, in which crosses are made to generate populations with known relatedness, has been the most commonly implemented method to identify QTLs responsible for phenotypic variation (Calenge et al., 2005; Dugo et al., 2005; Galliot et al., 2006; Messmer et al., 1999; Wang et al., 2011). In lily, linkage mapping has not been well established. So far, linkage mapping populations were established for Asiatic crosses 'AA' (Abe et al., 2002; Van Heusden et al., 2002) and for an inter-specific cross 'LA' (Khan, 2009). In this thesis, we constructed genetic maps for two populations: LA (*L. longiflorum* 'White Fox' x Asiatic hybrid 'Connecticut King') and AA ('Connecticut King' x 'Orlito') using AFLP (amplified fragment length polymorphisms), DArT (diversity arrays technology), NBS (nucleotide binding site) profiling, and SNP (single nucleotide polymorphisms) markers. The maps cover 89% and 74% of lily genome respectively with a resolution of one marker per 4 cM on average and as such should form a good basis for QTL studies and can act as a framework for further genetic studies.

Linkage mapping approach is a controlled method since relatedness between offspring is known, however only recombination events that have occurred during establishment of mapping population can be detected (Myles et al., 2009). This means that a QTL will be localized to a large chromosomal region (10-20 cM) depending on offspring number (Myles et al., 2009). Thus, the larger the population size and the more markers generated, the more accurate QTL regions will be defined. Also, linkage mapping can only identify genetic and phenotypic variations generated from their parents, which might often present only a small fraction of the variation in the whole species. This constraint limits the usability of linkage mapping.

Genome-wide association 'GWA' is proposed as an alternative for the traditional linkage mapping in several crops such as in grape and other long-lived perennial fruit crops (Myles et al., 2011). The GWA is usually defined as the detection of associations between molecular markers and phenotypic traits in a population of genotypes in which relatedness is not controlled (D'Hoop, 2009). Importantly, GWA exploits all recombination events that have occurred in the evolutionary history of the genotypes, which results in much higher mapping resolutions compared to linkage mapping (Myles et al., 2009). One of the very attractive aspects of GWA is that, it is not necessary to develop crosses and screen them for traits of interest. Instead, a collection of all genotypes (species, hybrids, cultivars, lines…) of the studied species can be genotyped and phenotyped for all interesting traits, not just the few that segregate from the parents as the case in linkage mapping. Applying GWA approach in lily can be interesting. A large collection of *Lilium* germplasm with huge genetic diversity that can serve as a source of different interesting traits for breeding purposes is available, thus there is no need to develop new linkage populations whenever the cultivars of interest are changed in markets. However, GWA approach is highly dependent on the ability to develop the necessary molecular markers required for mapping and gene identification. Also, one should be aware that if traits are present at low frequency in germplasm then they might be difficult to be spotted especially if the traits are influenced by environmental effects. Making a combination of the two strategies (linkage mapping and GWA) is particularly powerful (Schneeberger and Weigel, 2011). The QTL resolution in linkage mapping approach can be increased by combining mapping populations with GWA data. On the other side, linkage mapping is helpful for GWA to rule out false positive signals that occur due to the confounding effect of relatedness between individuals in germplasm used for GWA (Schneeberger and Weigel, 2011).

Developing genomic and genetic resources for *Lilium*

Next generation sequencing as a tool for resources development in *Lilium*

The recent emergence of second generation sequencing technologies like sequencing-by-synthesis (SBS) technologies, such as 454 Life Sciences (Roche Applied Science, Indianapolis, IN) and Illumina Genome Analyzer/Solexa (Illumina, San Diego, CA), or sequencing-by-ligation technology like ABI SOLiD (Life Technologies Corporation, Carlsbad, CA) (Deschamps and Campbell, 2010), and the very recent development of the third generation sequencing techniques such as: Pacific Biosciences (PacBio RS) based on a real-time, single-molecule sequencing technology, Complete Genomics based on a hybridization and ligation method, and Ion Torrent by Life Technologies based on sequencing by synthesis (Niedringhaus et al., 2011), have opened the door for large scale sequencing and thus providing huge amounts of sequence data beneficial for genetic and genomic studies (Mammadov et al., 2010).

The preference of one of these sequencing technologies is highly related on the purpose and the budget of the project. Sequence quality and read length, number of bases (Gbs to Tbs) produced per run, and the cost of each run varies among these technologies (Deschamps and Campbell,

2010). Although, the number of bases produced per run and the cost of each run are essential for choosing the sequencing technology, read length is a main player when dealing with genomes that lack support of sequence information from databases (Treangen and Salzberg, 2012). The 454 sequencing system is recommended for sequencing large genomes (Velasco et al., 2007; Wheeler et al., 2008), since it generates the longest read length (450 bp) compared to the other second generation technologies (Illumina (75-100 bp) and ABI SOLiD) which is essential for *de novo* assembly projects (Deschamps and Campbell, 2010; Paszkiewicz and Studholme, 2010). Longer reads are generally preferred, as they greatly reduce the complexity of assembly (Martin and Wang, 2011; Treangen and Salzberg, 2012). As for third generation sequencing technologies: PacBio produces the longest read lengths (up to 3,000 bases), however with low single-pass accuracy (81-83%) and high cost per base, Complete Genomics technology produces the highest (claimed) throughput of third generation platforms with short read length and very laborious sample preparation, and Ion Torrent produces short read length (100-200 pb) with a high potential for error accumulation and difficulties in sequencing highly repetitive or homopolymeric regions of the genome (Niedringhaus et al., 2011). Third generation sequencing technologies are not available commercially yet.

In our study, 454 pyro-sequencing was selected to sequence the transcriptome of *Lilium* since it was available technology at the time of conducting this research and produced the longest read lengths (average between 300 to 400 bp) compared to other second generation technologies. With the new developments in read length (100-150 bp), paired end sequence possibility and number of reads per run (six billion paired-end reads/run) at affordable costs, Illumina sequencing may have become a good alternative to 454 transcriptome sequencing and generate longer contigs and thus better assembly (Etter et al., 2011; Deschamps and Campbell, 2010).

Complexity reduction

A main challenge for sequencing is the huge and highly repetitive nature of some species genomes. From a computational perspective, repeats create ambiguities during assembly, which in turn produces errors when interpreting the results (Treangen and Salzberg, 2012). Thus the need for complexity reduction methodologies is essential when dealing with such a genome.

Complexity reduction methodologies that have been applied so far go under four categories: restriction enzymes (RE), degenerated primers, microarrays, and mRNA. Several techniques implemented RE to reduce complexity such as: reduced representation libraries (RRLs) in which genomic DNA is digested by RE followed by size selection (Hyten et al., 2010), complexity reduction of polymorphic sequences (CRoPS) using AFLP restriction ligation libraries for complexity reduction (Van Orsouw et al., 2007), restriction-site associated DNA (RAD-seq) tag sequencing (Emerson et al., 2010) in which genomic DNA is digested with a RE, then randomly sheared into very small fragments to isolate DNA tags directly flanking the restriction sites of a particular restriction enzyme throughout the genome (Miller et al., 2007), and methylation filtration by using methyl sensitive restriction enzymes to enrich for genes by targeting only the

hypo-methylated fraction of the genome (Whitelaw et al., 2003). Degenerated primers were applied to reduce genome complexity by the use of non-standard DOP primers, which contain G/C-rich 5' anchors (CTCGAG) followed by six degenerate nucleotides (N) and at least eight specific nucleotides at the 3' end (Janiak et al., 2008), or by the use of degenerated primers to amplify gene families (motif-directed profiling) such as resistance genes or peroxidase genes (González et al., 2010; Smulders et al., submitted; Van der Linden et al., 2005). It is also possible to use intron sequencing, in which intron-flanking EST-specific primers are used to amplify a relatively small number of loci which can be used to test large sets of genotypes more than getting lots of resources from one genotype (Wei et al., 2005). Microarray hybridization is based on selection of specific loci prior to sequencing by using a high-density oligonucleotide microarray to capture genome segments of interest (Albert et al., 2007), or hybridization to an identified set of single copy conserved orthologous (COS) genes which are shared by most, if not all, plant species (Li et al., 2008; Wu et al., 2006). Transcriptome sequencing (mRNA or RNA-seq), the complexity is reduced by sequencing only the transcriptomes isolated of certain tissue(s) (Deschamps and Campbell, 2010; Novaes et al., 2008).

In our study, RNA-sequencing was used to generate sequence resources (expressed sequence tags, EST) for four lily cultivars 'Connecticut King', 'While Fox', 'Star Gazer', and a 'Trumpet' hybrid, out of the four main lily groups: Asiatic (A), Longiflorum (L), Oriental (O), and Trumpet (T), respectively. EST sequence data are genetically informative (represent genes) and can be used for: molecular markers development, comparative genomics, gene annotation, and biodiversity and population genetic studies. Additionally, any polymorphism found will be in, or close to, genes. Genes are not the most polymorphic part of the genome, which might be considered (in general) a disadvantage (Smulders et al., submitted); however, in our study this might be preferred. *Lilium* is an outbreeding crop which means that the four genotypes are heterozygous, also these cultivars belong to four different hybrid groups and thus their genomes are expected to be highly heterogeneous when compared. Consequently, polymorphism rate is expected to be high. High polymorphism rates within sequence data sets complicate assembly process (Vinson et al., 2005). The assembly contiguity and completeness are significantly lower than would have been expected in the absence of heterozygosity, although factors such as: repeat content, segmental duplications, sequencing errors, homopolymer tracks, and library quality can also affect assembly quality (Vinson et al., 2005). Thus, reducing the number of detected SNPs by focusing only on polymorphisms present in genes is preferred in this case.

We used leaf tissue for mRNA extraction. As not all genes are being expressed in a given tissue and developmental stage, this might reduce the amount of sequences that could be generated. This can be captured by isolating mRNA of different tissues which are later pooled.

Sequence assembly

Post-sequencing computational treatment (trimming, cleaning, clonality removal, and assembly) might be the most important prerequisites for true representation of genes in studied genotypes

(Mammadov et al., 2010). There are three strategies for transcriptome assembly: reference-based strategy (applied when a reference genome for the target transcriptome is available), *de novo* strategy (applied when the reference genome for the target transcriptome is not available), and a combined strategy in which both strategies can be combined to create a more comprehensive transcriptome (Martin and Wang, 2011), where it is also possible to map reads to a closely related genome when it is available. In *Lilium* and *Tulipa*, *de novo* assembly was applied due to missing support information from databases.

Choosing the assembly program is very challenging (Martin and Wang, 2011), since the assembly is strongly depended upon the assembler (Feldmeyer et al., 2011). Two main approaches: the overlap layout consensus (OLC) and the De Bruijn were implemented in assembly programs. To avoid wrong assembly for our data, both approaches were tested. The CLC assembler (De Bruijn) showed to perform better than the CAP3 assembler (OLC), using assembly redundancy as a parameter to assess assembly quality, and thus CLC was used in our study. An important challenge for the assembler is to separate alleles from paralogs. If alleles and paralogs are not correctly assembled further studies, mainly SNP marker retrieval and comparative studies, will be complicated. In this study, transcriptome assembly (RNA-seq) of the four cultivars of *Lilium* resulted in 52,172 unigenes with an average read length of 555 bp. A similar approach was applied for *Tulipa* which resulted in identification of the first set of unigenes (81,791) for *Tulipa* with an average length of 514 bp.

Application of NGS for MAB: marker development

Molecular mapping through linkage mapping or GWA presents significant logistical challenges to develop markers that cover the majority if not all variation in the studied genome (Jordan et al., 2002). For accurate trait mapping and initiating MAB, good marker coverage of genetic maps is required, preferably with markers that target traits of interest and can be easily used in downstream breeding applications.

Traditionally, development of markers such as RAPD (random amplified polymorphic DNA), RFLPs (restriction fragment length polymorphisms), NBS profiling, and AFLPs is a costly, repetitive, and time-consuming process (Davey et al., 2011). Scoring of marker panels across target populations is also expensive, laborious, and not always objective (Davey et al., 2011).

By contrast, the invent of NGS technologies that generate huge amount of sequence data in a relatively very short time has enabled the discovery of thousands of SNP markers, and provided a resource for microsatellite marker identification. SNPs are considered to be handy tools for many genetic applications such as construction of genetic maps, discovery of QTLs, assessment of genetic diversity, pedigree verification, cultivar identification, association analysis, and marker-assisted breeding (Zhu et al., 2003). Moreover, availability of high throughput genotyping technology, which allows genotyping of thousands of SNP markers simultaneously in tens to

hundreds of individuals (Davey et al., 2011), and the objective (automated) scoring make SNP markers even more attractive.

SNP markers have a valuable role in MAB. Any SNP marker showing a link to a trait of interest can be used directly for downstream application in breeding. However, this is not the case for conventional marker systems. The AFLP, RAPD, and NBS markers need to be re-produced, cloned, and sequenced to track back the variation underlying the marker variance to develop a robust PCR marker for MAB application. Although, DArT markers do not need to be reproduced, they still need to be sequenced to track back the original variation and to develop a robust PCR marker based on this variation (Shahin et al., 2009). These conversion steps are very time consuming and challenging due to possible co-migrating bands and the highly repetitive sequences that restrict the conversion of AFLP and NBS bands into single locus markers. The conversion of DArT markers was successful for some DArT markers even though, it was complicated in some cases (Shahin et al., 2009).

Another important issue is the transferability of markers. The AFLP and NBS markers cannot be transferred from one parent to another, or between linkage mapping populations due to previously mentioned complication, while it was possible for DArT markers although laborious. This is clearer for SNP markers. When sequence data are available for the parents of a population, SNP presence can be verified in both parents, if not another SNP searched for from the same contig sequence (representing different polymorphism in the same gene) and used for mapping. This is very important for synteny and collinearity studies. *Lilium* germplasm is rich with all types of interspecific hybrids (*e.g.* AA, LL, OO, TT, LA, OT, OA, and LO). We intended in this study to develop molecular markers that are specific for each hybrid group, and markers among the four genotypes of lily (Chapter 4) to produce transferable markers that can be used for genotyping a wide genetic background and implemented in association and genetic diversity studies.

SNP markers genotyping

There are three general allele-discrimination methods implemented in genotyping technologies: hybridization/annealing (with or without a subsequent enzymatic step), primer extension and enzyme cleavage. Detection of allele-specific products are done by fluorescence detection, fluorescence resonance energy transfer (FRET), pyro-sequencing, mass spectrometry, or atomic force microscopy (AFM) (Ding and Jin, 2009; Twyman and Primrose, 2003).

In plants, the main SNP assays that have been developed until now are: TaqMan (single locus assay based on enzyme cleavage method, and the detection is done by FRET), KASP (KBiosciences competitive Allele Specific PCR, single locus assay which is a fluorescent FRET based system coupled with competitive allele specific PCR, http://www.kbioscience.co.uk/), GoldenGate (multiplex bead array 'Microarray', based on allele specific primer extension enabled by an enzymatic discrimination step that is visualized by fluorescence detection and

Infinium (multiplex bead assay 'Microarray', based on allele specific primer extension that is detected by fluorescence detection) (Ding and Jin, 2009).

The choice for a genotyping technology is strongly depending on the nature of the research. For example, to validate a small subset of SNP markers, TaqMan or KASP assays that use a flexible OpenArray platform can be best used. These technologies are high throughput but singleplex (Jiannis, 2006). To genotype high number of samples with high number of SNPs other technologies are preferred (Jiannis, 2006). The genotyping technology with the higher multiplexing abilities such as GoldenGate and Infinium assays are suitable for high-throughput genome-wide SNP development within a short period of time. These technologies allow multiplexing of up to 1,536 GoldenGate (Fan et al., 2003) and 200,000 SNPs Infinium assay, respectively, in one reaction within a 3-day period.

To improve genetic maps of *Lilium*, we applied KASP to genotype SNP markers derived from 454 pyro-sequencing of a cDNA library constructed of cultivar 'Connecticut King'. A total of 225 SNP markers were used for genotyping two mapping populations of lily (LA and AA populations). Genotyping success rate was 75.5% from which 45 % were polymorphic and the majority could be mapped. Development, data manipulating, and genotyping of SNP markers of transcriptome sequences generated using NGS technology of the uncharacterized lily genome were successful in AA and LA populations, and SNP markers showed a high efficiency in mapping. This opens the door for genotyping more SNP markers in these two populations. However, the usability of SNP markers for GWA in *Lilium* is not clear yet.

A main challenge for SNP marker genotyping is the SNPs in the flanking sequences of the target SNP which may have a significant impact on SNP genotyping performance (Grattapaglia et al., 2011). There is no systematic assessment of the effect of additional polymorphisms detected in the flanking sequence of the target SNP on its genotyping reliability. For species with a relatively low nucleotide diversity, such as humans and crop plants, this represents a minor concern; while it is a crucial issue for highly heterozygous genomes with nucleotide diversity in excess of 1% (Chancerel et al., 2011; Grattapaglia et al., 2011; Myles et al., 2011). In *Lilium*, the occurrence of null alleles in SNP markers developed from 'Connecticut King' was recorded in AA and LA populations (Chapter 5). Polymorphism rate among the four sequenced *Lilium* cultivars was rather high (one SNP per 26 bps, 4%), which might be the cause of null-alleles occurrence and consequently complicates SNP genotyping on wide collections of germplasm. This might be improved by sequencing more genotypes (mainly those widely used as parents in breeding programs) for sufficient depth and comparing their sequences to select the most conserved regions for SNP markers development. Alternatively, genotyping by sequencing (GBS) using Myselect (http://www.mycroarray.com/myselect/MYselect+sequence+capture+target+enrichment+kit.html) technology might give an appreciable solution. Myselect technology uses ESTs information to target genes of interest by using sequence capture bait libraries constructed by synthesizing large long-oligonucleotide based on EST sequences. Applying such technology will help to reduce

genotyping failures due to additional SNPs present in the flanking sequences of a SNP marker while additionally picking up yet undiscovered other SNPs.

GBS is becoming more feasible due to the advances in NGS that have driven the costs of DNA sequencing down (Elshire et al., 2011). GBS can be performed in combination with a number of complexity reduction methods like RAD followed by NGS technology (Elshire et al., 2011) or by using baits developed from EST sequences (Myselect technology, explained previously) depending on the target of the research. While RAD might be preferred for SNP markers development, RNA-seq could be preferred for resource development projects. The combination of RAD and GBS is technically simple and highly amendable for multiplexing, and can be used for a variety of applications like population studies, germplasm characterization, breeding, and trait mapping in diverse organisms (Elshire et al., 2011). However, for highly heterogeneous species, the combination of Myselect technology with GBS might be preferred to avoid the null allele problem expected when using SNP markers.

Application of NGS for comparative studies

Synteny between *Lilium* and *Tulipa*

Exchanging genetic information between two related species by direct comparison of DNA sequences and map positions will contribute to the understanding of their evolution and divergence mechanisms. To date, relations, at the genome sequence level, have been demonstrated within several families: *Poaceae*, *Solanaceae*, *Brassicacea*, *Fabaceae*, and *Pinus*, as well as between families (*e.g.* *Arabidopsis*/rice, *Arabidopsis*/tomato, and *Arabidopsis*/moss) (Bennetzen and Ma, 2003; Chancerel et al., 2011; Erpelding et al., 1996; Fulton et al., 2002; Izawa et al., 2003; Ku et al., 2000; Lagercrantz and Lydiate, 1996; Lan et al., 2000; Livingstone et al., 1999; Nishiyama et al., 2003; Park et al., 2011; Salse et al., 2002; Thorup et al., 2000; Verlaan et al., 2011). Using synteny, mapping the evolutionary events like: duplications, rearrangements, and conserved regions can be detected (Tang et al., 2011). Also, micro-synteny (homeologous region) allows identification of a gene in one plant species based on detailed positional and sequence information in the homeologous region of another genus. However this approach is not straightforward due to genome rearrangement events and differences in evolutionary rates for different lineages and gene families (McCouch, 2001; Tang et al., 2011).

Transferability of genetic information between related species has been studied by several techniques such as: sequence tagged site primers developed from mapped RFLP clones to link genetic maps of wheat and barley (Erpelding et al., 1996) and tropical grains with grasses (Bowers et al., 2003); selection of highly conserved ESTs such as between *Arabidopsis thaliana* and *Brassica oleracea* (Lan et al., 2000), between *Oryza sativa* and wheat (Rota and Sorrells, 2004); microarray for transcriptional comparative in *Solanaceae* (Moore et al., 2005) and BAC sequences coupled with FISH (fluorescence *in situ* hybridization) in *Solanaceae* family (Park et al., 2011; Verlaan et al., 2011). Recently, several studies implemented SNP data for comparative

studies in large and un-sequenced genomes. For instances: GoldenGate assay was used (SNPs detected from NGS and Sanger sequences) to link the linkage maps of *Pinus pinaster* and *Pinus taeda* (Chancerel et al., 2011), *Picea mariana* and *Picea glauca* (Pavy et al., 2008), *Eucalyptus grandis*, and *Arabidopsis thaliana* (Novaes et al., 2008), and to identify conserved synteny between *Jatropha* and castor bean, popular, and *Arabidopsis* (Wang et al., 2011).

In *Liliaceae* family, so far, comparative genome analysis to link genetic maps or physical maps has not been performed yet. We initiated the first step towards linking molecular genetic maps of *Lilium* and *Tulipa* using transcriptome sequences generated by NGS technology. Orthologous genes between lily and tulip (10,913 unigenes) were identified based on sequence data of four lily cultivars and five tulip cultivars. Next, common SNP and EST-SSR markers between the parents of lily mapping populations (AA and LA populations) and the parents of the mapping population of *Tulipa* ('Kees Nelis' (*T. gesneriana*) x 'Cantata' (*T. fosteriana*)) based on orthologous sequences of both species were identified: 229 common SNP (not present the same polymorphism), and 140 common EST-SSR. Genotyping these markers in both populations will initiate the link between the genetic maps of *Lilium* and *Tulipa*. The efficiency of these markers in comparative study depends largely upon how many of these markers will be mapped on both the lily and tulip genetic maps and also on how well these markers are distributed over the chromosomes. Mapping and comparing the common markers allow insight into the preservation of gene order, structure, and 'putative' functional homology in addition to evolution and divergence mechanisms (Moore et al., 2005). The Gene Ontology (GO) assessment of the 10,913 orthologous sequences showed that majority of them represent genes essential for growing and defense processes, in addition to genes related to biological processes. Overall, the majority of orthologous genes were housekeeping genes.

Systematics of *Lilium* and *Tulipa*

Increasing accessibility and affordability of NGS provides large amounts of data from the three organelles of plants (chloroplast, mitochondrion, and nucleus) in a single run, which can be used to increase the resolution of, and support for, phylogenetic trees (Steele and Pires, 2011; Straub et al., 2012). In this thesis, we could identify a set of orthologous genes of *Lilium* (19 genes), of *Tulipa* (20 genes), and of orthologous genes between the two *genera* (7 genes), which showed to be uniquely present in the sequences and informative in estimating the genetic divergence of these two genera, thus they can be used to study the phylogeny of *Lilium* and *Tulipa* in depth. This can be done by developing probes for each gene, attach them to baits and use them to fish for this sequences in a set of genotypes, followed by sequencing these genes using NGS (Myselect approach). To build a genus tree for *Lilium* or *Tulipa*, several species per section in additions to hybrids among sections should be used to cover the genetic diversity present in each genus. By sequencing those genes in several species per genus, gene trees and genus tree can be constructed. This kind of vast sequence data will be helpful to resolve discrepancies, if present, between genetic and morphological classifications such as in *Lilium* (Chapter 6).

Sequence concatenation is no longer an issue since new software such as BEST (Liu et al., 2008) and BEAST* (Drummond and Rambaut, 2007; Heled and Drummond, 2010) were introduced to estimate species trees from gene trees and to deal with the multi-allelic nature of genes. However, using consensus sequences in which SNPs between alleles of a gene become ambiguous (IUPAC bases), leads to loss of part of the available data. Another possibility is to use the average genetic distances between the haplotypes, POFAD (Phylogeny of Organisms from Allelic Data) algorithm (Joly and Bruneau, 2006), however this methodology does not provide means to test statistical significance by bootstrapping, and it showed that the resolution was reduced compared with a consensus sequence approach (Chapter 6). Nevertheless, using allelic information is interesting mainly in revealing phylogenetic relations between closely related taxa and in species hybrids. Thus, further studies are needed to find the best methodology that can implement allelic variation. Apart from that, a pipeline that includes both GBS data analysis and formatting for tree building software like BEAST* can be developed which will help to automate the whole process starting from genotyping, gene trees, and species tree construction.

Having such high numbers of sequence data, allows us to test some evolutionary hypotheses. One interesting hypothesis is that breeding or human selection processes might be imprinted on genome. Such hypotheses can be studied by comparing cultivar genomes with wild species genomes. Another one is whether *Lilium* and *Tulipa* have undergone bottlenecks during their domestication process or not. In outbred species, such as grape, a weak bottleneck has been recorded (Myles et al., 2011), while in cereal crops (inbred) such as wheat, rice, and maize (Buckler et al., 2001; Eyre-Walker et al., 1998; Zhu et al., 2007) strong bottlenecks are observed. Weak bottlenecks are expected in *Lilium* and *Tulipa* since they are both outbred and vegetative propagated similar to grape.

Agronomical traits in *Lilium*

Mapping of disease resistance

Fusarium oxysporum and lily mottle virus (LMoV) are considered as very serious diseases in *Lilium* (Shahin et al., 2009; Straathof et al., 1996; Van Heusden et al., 2002). *Fusarium oxysporum* was mapped on the genetic map of AA population, in which six putative QTLs were identified (not all of them confirmed by interval mapping test). QTL1 is a strong QTL that showed to be linked to *Fusarium* resistance in several years of disease testing. However the resolution of this QTL is still low and more markers are needed to increase the mapping resolution. Moreover, mapping *Fusarium* in LA population has been also performed using disease test results of two years. The mapping was resulted in a strong QTL that is co-localizing with the QTL1 of AA population (data not shown). This strong and reliable QTL identified in both populations provides a good start for generating markers for *Fusarium* resistance for MAB applications. Genotyping and mapping some SNP markers in these populations did not result in providing SNP markers linked to this QTL yet (data not shown).

LMoV was mapped as a marker on AA genetic maps; however no close markers have been identified. By adding SNP markers to AA genetic maps, LMoV locus was changed and showed to be closer to some markers (data not shown) that can be used for validation and MAB.

Mapping of ornamental traits

There are several important traits that ornamental breeders wish to improve such as: diseases resistances (*Fusarium*, *Botrytis* and viruses), postharvest quality, flower color, and scent. In our study, in which the genetic maps of two populations became available, we mapped in *Lilium* several ornamental traits. These traits are: lily flower color 'carotene' '*LFCc'*, flower spots '*lfs'*, stem color '*LSC'*, antherless phenotype '*lal*', and flower direction '*lfd'*. Several of them showed to be recessive traits (spots, antherless, and flower direction). For breeding, availability of markers for recessive ornamental traits is very useful. Such markers allow the identification of suitable breeding parents so that expression of the recessive trait can be either enhanced or repressed.

Identification of QTLs for complex traits like vase life is rather complicated and very time consuming. Several studies investigated the possibilities for improving flower longevity of Asiatic hybrid lilies by breeding (Lim, 2000; Van der Meulen-Muisers et al., 1999; Van der Meulen-Muisers et al., 1998; Van der Meulen et al., 1996; Van der Meulen et al., 1997). Van der Meulen et al (1996) screened two Asiatic populations: 'Yellito' x 'Orlito' and 'Connecticut King' x 'Orlito' for their longevity and compared their morphological data with molecular markers and identified few markers (RAPD) linked to longevity. However, the limited number of genotyped progeny (39 genotyped), the lacking of genetic maps, and thus the possibility of applying only the non-parametric test of Kruskal-Wallis to find linkages, and the testing of only one bulb per genotype for longevity have limited the application of this study (Van der Meulen et al., 1996). Also, the possibility of identifying undesirable non-genetic variation resulting from growing conditions, development stage of the flower at harvest and environmental conditions after harvest (Van der Meulen-Muisers et al., 1998) make this trait very challenging for studying. In addition, longevity in lilies is complex because it is a function of: the number of buds per inflorescence, the expansion and opening of buds, the lifespan of individual flowers, and the lifespan of leaves (Van der Meulen-Muisers et al., 1999). Thus, an alternative approach should first be applied to understand and simplify this trait, followed by identifying markers associated with this trait.

Vase life

In *Lilium*, even though senescence is ethylene-insensitive, lily inflorescences from AA and LA cultivars have to be treated with an ethylene-inhibitor (silver thiosulfate) before they are sent to auctions. The effect of ethylene-inhibitor on *Lilium* vase life is doubtful (Elgar et al., 1999; Nowak and Mynett, 1985; Van der Meulen-Muisers et al., 2001), and the regulators of lily flower longevity are not known yet (Chapter 7). We aimed to define the regulator(s) of flower longevity in lily by studying the effect of sugar (sucrose) addition on flower longevity and abscisic acid (ABA) hormonal changes between anthesis and senescence.

Vase life of lily flowers increased significantly by the exogenous application of sugars, and a dramatic increase in ABA levels at senescence compared with anthesis was recorded. This might indicate ABA as a main regulatory factor in controlling vase life of lily and the potential vital role of sugar in delaying all senescence symptoms. Interestingly, sugar treatment decreased ABA concentration, suggesting an opposite effects of ABA and sugar on lily senescence and a possible crosstalk between the pathways of both. It is important to understand the different mechanisms that control vase life in lily which will have a direct implication on the ornamental and thus commercial value of this flower crop.

To confirm the role of ABA, an external application of ABA inhibitor should be applied to lily inflorescences to study its effect on longevity. This could be combined with transcriptional comparative analysis (RNA-seq) using lily flowers at anthesis and senescence to determine genes that are up or down regulated. Several proteins have been reported to function as ABA receptors and many more are known to be involved in ABA signaling (Finkelstein et al., 2002; Schroeder et al., 2001; Shen et al., 2006). By analyzing RNA-seq data, genes of the ABA pathway that show to be associated with senescence can be determined.

Overall, in this study, we generated and characterized the transcriptome of four *Lilium* genotypes which have direct applications for breeding in this species. The use of NGS technologies provides a wealth of putative molecular markers (SNPs and SSRs) that can be used for genotyping the two mapping lily populations used in this study. This will improve both the coverage of the *Lilium* genome and the marker density of the two populations. Also, genotyping the same SNP marker in the two populations will facilitate the comparisons between linkage group populations and allow construction of a consensus map. Consequently, exchange of genetic knowledge (mainly QTLs) between the two populations will be easier. Additionally, the thousands of SNPs identified in genome of four lily cultivars opens the door for combining current linkage mapping studies with association studies which will have a direct impact on improving the resolution of mapping and on marker assisted breeding in *Lilium*.

References

Abe, H., Nakano, M., Nakatsuka, A., Nakayama, M., Koshioka, M., and Yamagishi, M. (2002). Genetic analysis of floral anthocyanin pigmentation traits in Asiatic hybrid lily using molecular linkage maps. Theor. Appl. Genet. 105:1175-1182.

Agrama, H.A.S., and Moussa, M.E. (1996). Mapping QTLs in breeding for drought tolerance in maize *(Zea mays* L.). Euphytica 91:89-97.

Akbari, M., Wenzl, P., Caig, V., Carling, J., Xia, L., Yang, S., Uszynski, G., Mohler, V., Lehmensiek, A., Kuchel, H., Hayden, M., Howes, N., Sharp, P., Vaughan, P., Rathnell, B., Huttner, E., and Kilian, A. (2006). Diversity arrays technology (DArT) for high-throughput profiling of the hexaploid wheat genome. Theor. Appl. Genet. 113:1409-1420.

Akutsu, M., Ishizaki, T., and Sato, H. (2004). Transformation of the monocot *Alstroemeria* by *Agrobacterium rhizogenes*. Mol. Breed. 13:69-78.

Alavi-Kia, S.S., Mohammadi, S.A., Aharizad, S., and Moghaddam, M. (2008). Analysis of genetic diversity and phylogenetic relationships in *Crocus* genus of Iran using inter-retrotransposon amplified polymorphism. Biotechnol & Biotechnol Eq. 22:795-800.

Albert, T.J., Molla, M.N., Muzny, D.M., Nazareth, L., Wheeler, D., Song, X., Richmond, T.A., Middle, C.M., Rodesch, M.J., Packard, C.J., Weinstock, G.M.and Gibbs, R.A. (2007). Direct selection of human genomic loci by microarray hybridization. Nat. Methods 4:903-905.

Albini, S.M., and Jones, G.H. (1990). Synaptonemal complex spreading in *Allium cepa* and *Allium fistulosum*. III. The F1 hybrid. Genome 33:854-866.

Alvarez, I., and Wendel, J.F. (2003). Ribosomal ITS sequences and plant phylogenetic inference. Mol. Phylogenet. Evol. 29:417-434.

Aneja, M., Gianfagna, T., and Ng, E. (1999). The roles of abscisic acid and ethylene in the abscission and senescence of cocoa flowers. Plant Growth Regul. 27:149-155.

Anisimova, M., Bielawski, J.P., and Yang, Z. (2001). Accuracy and Power of the Likelihood Ratio Test in Detecting Adaptive Molecular Evolution. Mol. Biol. Evol. 18:1585-1592.

Anisimova, M., Nielsen, R., and Yang, Z. (2003). Effect of Recombination on the Accuracy of the Likelihood Method for Detecting Positive Selection at Amino Acid Sites. Genetics 164:1229-1236.

Anithakumari, A.M., Tang, J., Van Eck, H.J., Visser, R.G., Leunissen, J.A., Vosman, B., and Van der Linden, C.G. (2010). A pipeline for high throughput detection and mapping of SNPs from EST databases. Mol. Breed. 26:65-75.

Arenas-Huertero, F., Arroyo, A., Zhou, L., Sheen, J., and León, P. (2000). Analysis of *Arabidopsis* glucose insensitive mutants, gin5 and gin6, reveals a central role of the plant hormone ABA in the regulation of plant vegetative development by sugar. Genes Dev. 14:2085-2096.

Aros, D., Meneses, C., and Infante, R. (2006). Genetic diversity of wild species and cultivated varieties of alstroemeria estimated through morphological descriptors and RAPD markers. Sci. Hortic. 108:86-90.

Arzate-Fernandez, A.M., Mejía-González, C.O., Nakazaki, T., Okumoto, Y., and Tanisaka, T. (2005). Isozyme electrophoretic characterization of 29 related cultivars of lily (*Lilium* spp.). Plant Breed. 124:71-78.

Asano, Y. (1980). Studies on crosses between distantly related species of Lilies. V. Characteristics of newly obtained hybrids through embryo culture. J. Japan. Soc. Hortic. Sci. 49:241 – 250.

Asano, Y. (1989). *Lilium*. In: The grand dictionary of horticulture--Tsukamoto, Y., ed.: Shogakukon, Tokyo, Japan. 198-209.

Asano, Y., and Myodo, H. (1977b). Studies on crosses between distantly related species of Lilies. II. The culture of immature hybrid embryos. J. Japan. Soc. Hortic. Sci. 46:267 - 273

Asano, Y., and Myodo, H. (1977a). Studies on crosses between distantly related species of lilies I. For the intrastylar pollination technique. J. Japan. Soc. Hort. Sci. 46:59-65

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S.,

Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genet. 25:25-29.

Azad, A.K., Ishikawa, T., Ishikawa, T., Sawa, Y., and Shibata, H. (2008). Intracellular energy depletion triggers programmed cell death during petal senescence in tulip. J. Exp. Bot. 59:2085-2095.

Bach Holm, P. (1976). The C and Q banding patterns of the chromosomes of *Lilium longiflorum* (thunb.). Carlsberg Research Communications 41:217-224.

Ballarin, A., Prisa, D., Grassotti, A., Burchi, G. and Pierandrei, F. (2009). Leaf treatments to improve cut flower quality in easter lily and in Asiatic and Oriental hybrid lily cultivars. Acta Hort. 847:369-376.

Barba-Gonzalez, R., Lokker, A.C., Lim, K.B., Ramanna, M.S., and Van Tuyl, J.M. (2004). Use of 2n gametes for the production of sexual polyploids from sterile Oriental × Asiatic hybrids of lilies (*Lilium*). Theor. Appl. Genet. 109:1125-1132.

Barba-Gonzalez, R., Miller, C.T., Ramanna, M.S., and Van Tuyl, J.M. (2006). Induction of 2n gametes for overcoming F1-sterility in lily and tulip. Acta Hort. 714:99-106.

Barba-Gonzalez, R., Ramanna, M.S., Visser, R.G.F., and Van Tuyl, J.M. (2005). Intergenomic recombination in F1 lily hybrids (*Lilium*) and its significance for genetic variation in the BC1 progenies as revealed by GISH and FISH. Genome 48:884-894.

Barrier, M., Bustamante, C.D., Yu, J., and Purugganan, M.D. (2003). Selection on Rapidly Evolving Proteins in the *Arabidopsis* Genome. Genetics 163:723-733.

Batley, J., Barker, G., O'Sullivan, H., Edwards, K.J., and Edwards, D. (2003). Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. Plant Physiol. 132:84-91.

Beiki, A.H., Keifi, F., and Mozafari, J. (2010). Genetic differentiation of *Crucus* species by random amplified polymorphic DNA. GEB J.18.

Beninda-Emonds, O.R.P. (2004). Phylogenetic Super Trees: Combining Information to Reveal the Tree of Life: Springer Verlag, New York- ISBN 1402023286, pp550.

Bennetzen, J.L., and Ma, J. (2003). The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis. Curr. Opin. Plant Biol. 6:128-133.

Benschop, M., Kamenetsky, R., Le Nard, M., Okubo, H., and De Hertogh, A. (2010). The Global Flower Bulb Industry: Production, Utilization, Research. In: Horticultural Reviews: John Wiley & Sons, Inc. 1-115.

Bieleski, R.L. (1993). Fructan hydrolysis drives petal expansion in the ephemeral daylily flower. Plant Physiol. 103:213-219.

Blanca, J., Canizares, J., Roig, C., Ziarsolo, P., Nuez, F., and Pico, B. (2011). Transcriptome characterization and high throughput SSRs and SNPs discovery in *Cucurbita pepo* (*Cucurbitaceae*). BMC Genomics 12:104.

Blanco, A., Bellomo, M.P., Cenci, A., De Giovanni, C., D'Ovidio, R., Iacono, E., Laddomada, B., Pagnotta, M.A., Porceddu, E., Sciancalepore, A., Simeone, R., and Tanzarella, O.A. (1998). A genetic linkage map of durum wheat. Theore. Appl. Genet. 97:721-728.

Booy, G., Schoot, J., and Vosman, B. (2000). Heterogeneity of the internal transcribed spacer 1 (ITS1) *in Tulipa* (*Liliaceae*). Plant Syst. Evol. 225:29-41.

Borochov, A., Mayak, S., and Halevy, A.H. (1976). Combined Effects of Abscisic Acid and Sucrose on Growth and Senescence of Rose Flowers. Physiol. Plant 36:221-224.

Bouck, A., Peeler, R., Arnold, M.L., and Wessler, S.R. (2005). Genetic Mapping of Species Boundaries in Louisiana Irises Using IRRE Retrotransposon Display Markers. Genetics 171:1289-1303.

Bouck, A., Wessler, S.R., and Arnold, M.L. (2007). QTL Analysis of Floral Traits in Louisiana Iris Hybrids. Evolution 61:2308-2319.

Bowers, J., Abbey, C., Anderson, S., Chang, C., Draye, X., Hoppe, A., Jessup, R., Lemke, C., Lennington, J., Li, Z., Lin, Y., Liu, S., Luo, L., Marler, B., Ming, R., Mitchell, S., Qiang, D., Reischmann, K., Schulze, S., Skinner, D., Wang, Y., Kresovich, S., Schertz, K., and Paterson, A. (2003). A high-

density genetic recombination map of sequence-tagged sites for Sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses. Genetics 165:367 - 386.

Bräutigam, A., Mullick, T., Schliesky, S., and Weber, A.P.M. (2011). Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C3 and C4 species. J. Exp. Bot. 62:3093-3102.

Bryan, E. (2002). Bulbs. Timber Press, Inc, Portland, Oregon, ISBN-10: 0881925292, pp:454-462.

Bryan, J.E. (1989). Bulbs: Timber Press. ISBN-10: 0881921017, pp 750.

Bryant, D., and Moulton, V. (2004). Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks. Mol. Biol. Evol. 21:255-265.

Buckler, E.S., Thornsberry, J.M., and Kresovich, S. (2001). Molecular diversity, structure and domestication of grasses. Genet. Res. Camb. 77:213-218.

Buerkle, C.A., Gompert, Z., and Parchman, T.L. (2011). The n=1 constraint in population genomics. Mol. Ecol. 20:1575-1581.

Burchi, G., Nesi, B., Grassoti, A., Mensuali-Sodi, A., Ferrante, A. (2005). Longevity and ethylene production during development stages of two cultivars of *Lilium* flowers ageing on plant or in vase. Acta Hort. 682:813-822.

Burchi, G., Prisa, D., Ballarin, A. and Grassotti. (2011). Effect of leaf treatments on flower quality and shelf life in Asiatic lily. Acta Hort. 900:19-24.

Calenge, F., Van der Linden, C.G., Van de Weg, E., Schouten, H.J., Van Arkel, G., Denancé, C., and Durel, C.E. (2005). Resistance gene analogues identified through the NBS-profiling method map close to major genes and QTL for disease resistance in apple. Theor. Appl. Genet. 110:660-668.

Chancerel, E., Lepoittevin, C., Le Provost, G., Lin, Y.-C., Jaramillo-Correa, J., Eckert, A., Wegrzyn, J., Zelenika, D., Boland, A., Frigerio, J.-M., Chaumeil, P., Garnier-Gere, P., Boury, C., Grivet, D., Gonzalez-Martinez, S., Rouze, P., Van de Peer, Y., Neale, D., Cervera, M., Kremer, A., and Plomion, C. (2011). Development and implementation of a highly-multiplexed SNP array for genetic mapping in maritime pine and comparative mapping with loblolly pine. BMC Genomics 12:368.

Chandler, S. (2007). Practical lessons in the commercialization of genetically modified plants-Long vase-life carnation. Acta Hort. 764:71-81.

Chandler, S., and Tanaka, Y. (2007). Genetic Modification in Floriculture. Crit. Rev. Plant Sci. 26:169-197.

Chase, M.W., and Reveal, J.L. (2009). A phylogenetic classification of the land plants to accompany APG III. Bot. J. Linnean Soc. 161:122-127.

Chauvin, J.E., Hamann, H., Cohat, J., and Le Nard, M. (1997). Selective agents and marker genes for use in genetic transformation of *Gladiolus grandiflorus* and *Tulipa gesneriana*. Acta Hort. 430:291-298.

Chen, Q., Han, Z., Jiang, H., Tian, D., and Yang, S. (2010). Strong Positive Selection Drives Rapid Diversification of *R* Genes in *Arabidopsis* Relatives. J. Mol. Evol. 70:137-148.

Cheng, F.S., Weeden, N.F., and Brown, S.K. (1996). Identification of co-dominant RAPD markers tightly linked to fruit skin color in apple. Theor. Appl. Genet. 93:222-227.

Cheng, W.H., Endo, A., Zhou, L., Penney, J., Chen, H.-C, Arroyo, A., Leon, P., Nambara, E., Asami, T., Seo, M., Koshiba, T., and Sheen, J. (2002). A Unique Short-Chain Dehydrogenase/Reductase in *Arabidopsis* Glucose Signaling and Abscisic Acid Biosynthesis and Functions. Plant Cell 14:2723-2743.

Cheung, F., Haas, B., Goldberg, S., May, G., Xiao, Y., and Town, C. (2006). Sequencing *Medicago truncatula* expressed sequenced tags using 454 Life Sciences technology. BMC Genomics 7:272.

Cho, W.K., and Kim, J.Y. (2009). Integrated analyses of the rice secretome. Plant Signal. Behav. 4:345-347.

Churchill, G.A., and Doerge, R.W. (1994). Empirical Threshold Values for Quantitative Trait Mapping. Genetics 138:963-971.

Comber, H.F. (1949). A new classification of the genus *Lilium*. Lily Yearbook, Royal Hort. Soc.13:86-105.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674-3676.

Cork, J.M., and Purugganan, M.D. (2005). High-Diversity Genes in the *Arabidopsis* Genome. Genetics 170:1897-1911.

Custers, J.B., Eikelboom, W., Bergervoet, J.H., and Van Eijk, J.P. (1995). Embryo-rescue in the genus *Tulipa* L.; Successful direct transfer of *T.kaufmanniana* Regel germplasm into *T. gesneriana* L. Euphytica 82:253-261.

D'Hoop, B.B. (2009). Association mapping in tetraploid potato: Wageningen Universiteit. ISBN 9789085853336, pp 161.

Davey, J.W., Hohenlohe, P.A., Etter, P.D., Boone, J.Q., Catchen, J.M., and Blaxter, M.L. (2011). Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Rev. Genet. 12:499-510.

De Jong, P.C. (1974). Some notes on the evolution of lilies. The Lily Yearbook of the North American Lily Society. 27:23-28.

De la Torre, J., Egan, M., Katari, M., Brenner, E., Stevenson, D., Coruzzi, G.and DeSalle, R. (2006). ESTimating plant phylogeny: lessons from partitioning. BMC Evol. Biol. 6:48.

Debener, T., and Mattiesch, L. (1999). Construction of a genetic linkage map for roses using RAPD and AFLP markers. Theor. Appl. Genet. 99:891-899.

Deschamps, S.p., and Campbell, M. (2010). Utilization of next-generation sequencing platforms in plant genomics and genetic variant discovery. Mol. Breed. 25:553-570-570.

Ding, C., and Jin, S. (2009). High-throughput methods for SNP genotyping. Methods Mol. Biol. 578:245-254.

Doorenbos, J. (1954). Notes on the history of bulb breeding in the Netherlands. Euphytica 3:1-18.

Drummond, A., and Rambaut, A. (2007). BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7:214.

Dubey, A., Farmer, A., Schlueter, J., Cannon, S.B., Abernathy, B., Tuteja, R., Woodward, J., Shah, T., Mulasmanovic, B., Kudapa, H., Raju, N.L., Gothalwal, R., Pande, S., Xiao, Y., Town, C.D., Singh, N.K., May, G.D., Jackson, S., and Varshney, R.K. (2011). Defining the transcriptome assembly and its use for genome dynamics and transcriptome profiling studies in Pigeonpea (*Cajanus cajan* L.). DNA Res. 18:153-164.

Dubouzet, J.G., and Shinoda, K. (1999). Phylogenetic analysis of the internal transcribed spacer region of Japanese *Lilium* species. Theor. Appl. Genet. 98:954-960.

Dugo, M.L., Satovic, Z., Millán, T., Cubero, J.I., Rubiales, D., Cabrera, A., and Torres, A.M. (2005). Genetic mapping of QTLs controlling horticultural traits in diploid roses. Theor. Appl. Genet. 111:511-520.

Dunemann, F., Kahnau, R., and Stange, I. (1999). Analysis of complex leaf and flower characters in Rhododendron using a molecular linkage map. Theor. Appl. Genet. 98:1146-1155.

Dutta, S., Kumawat, G., Singh, B., Gupta, D., Singh, S., Dogra, V., Gaikwad, K., Sharma, T., Raje, R., Bandhopadhya, T., Datta, S., Singh, M., Bashasab, F., Kulwal, P., Wanjari, K., Varshney, R. K, Cook, D., and Singh, N., (2011). Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [*Cajanus cajan* (L.) Millspaugh]. BMC Plant Biol. 11:17.

Eason, J.R., Ryan, D.J., Pinkney, T.T., and O'Donoghue, E.M. (2002). Programmed cell death during flower senescence: isolation and characterization of cysteine proteinases from *Sandersonia aurantiaca*. Funct. Plant Biol. 29:1055-1064.

Elgar, H.J., Woolf, A.B., and Bieleski, R.L. (1999). Ethylene production by three lily species and their response to ethylene exposure. Postharvest Biol. Technol. 16:257-267.

Elshire, R.J., Glaubitz, J.C., Sun, Q., Poland, J.A., Kawamoto, K., Buckler, E.S., and Mitchell, S.E. (2011). A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species. PLoS ONE 6:e19379.

Emerson, K.J., Merz, C.R., Catchen, J.M., Hohenlohe, P.A., Cresko, W.A., Bradshaw, W.E., and Holzapfel, C.M. (2010). Resolving postglacial phylogeography using high-throughput sequencing. Proc. Natl Acad. Sci. USA 107:16196-16200.

Emrich, S.J., Aluru, S., Fu, Y., Wen, T.-J., Narayanan, M., Guo, L., Ashlock, D.A., and Schnable, P.S. (2004). A strategy for assembling the maize (*Zea mays* L.) genome. Bioinformatics 20:140-147.

Erpelding, J.E., Blake, N.K., Blake, T.K., and Talbert, L.E. (1996). Transfer of sequence tagged site PCR markers between wheat and barley. Genome 39:802-810.

Etter, P.D., Preston, J.L., Bassham, S., Cresko, W.A., and Johnson, E.A. (2011). Local *De Novo* Assembly of RAD Paired-End Contigs Using Short Sequencing Reads. PLoS ONE 6:e18561.

Eyre-Walker, A., Gaut, R.L., Hilton, H., Feldman, D.L., and Gaut, B.S. (1998). Investigation of the bottleneck leading to the domestication of maize. Proc. Natl Acad. Sci. USA 95:4441-4446.

Fan, J.-B., A. Oliphant, R. Shen, B.G. Kermani, F. Garcia, K.L. Gunderson, M. Hansen, F. Steemers, S.L. Butler, P. Deloukas, L. Galver, S. Hunt, C. Mcbride, M. Bibikova, T. Rubano, J. Chen, E. Wickham, D. Doucet, W. Chang, D. Campbell, B. Zhang, S. Kruglyak, D. Bentley, J. Haas, P. Rigault, L. Zhou, J. Stuelpnagel, and M.S. Chee (2003). Highly Parallel SNP Genotyping. Cold Spring Harbor Symposia on Quantitative Biology 68:69-78.

Fauré, S., Noyer, J.L., Horry, J.P., Bakry, F., Lanaud, C., and Gońzalez de León, D. (1993). A molecular marker-based linkage map of diploid bananas (*Musa acuminata*). Theor. Appl. Genet. 87:517-26.

Fay, M.F., Chase, M.W., Rønsted, N., Devey, D.S., Pillon, Y., Pires, J.C., Petersen G., Seberg, O. and Davis, J.I. (2006). Phylogenetics of Liliales: summarized evidence from combined analyses of five plastid and one mitochondrial loci. Aliso 22:559 -565

Feldmeyer, B., Wheat, C., Krezdorn, N., Rotter, B., and Pfenninger, M. (2011). Short read Illumina data for the de novo assembly of a non-model snail species transcriptome (Radix balthica, Basommatophora, Pulmonata), and a comparison of assembler performance. BMC Genomics 12:317.

Ferrante, A., Vernieri, P., Tognoni, F., and Serra, G. (2006). Changes in abscisic acid and flower pigments during floral senescence of petunia. Biol. Plant. 50:581-585.

Finkelstein, R., Gampala, S., and Rock, C. (2002). Abscisic acid signaling in seeds and seedlings. Plant Cell 14 Suppl:S15-45.

Finkelstein, R.R., and Gibson, S.I. (2002). ABA and sugar interactions regulating development: cross-talk or voices in a crowd? Curr. Opin. Plant Biol. 5:26-32.

Fitzpatrick, D., Logue, M., Stajich, J., and Butler, G. (2006). A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. 6:99.

Foolad, M.R., Arulsekar, S., Becerra, V., and Bliss, F.A. (1995). A genetic map of Prunus based on an interspecific cross between peach and almond. Theor. Appl. Genet. 91:262-269.

Franssen, S.U., Shrestha, R.P., Bräutigam, A., Bornberg-Bauer, E., and Weber, A.P.M. (2011). Comprehensive transcriptome analysis of the highly complex *Pisum sativum* genome using next generation sequencing. BMC Genomics 12:227.

Fulton, T.M., Van der Hoeven, R., Eannetta, N.T., and Tanksley, S.D. (2002). Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell 14:1457-1467.

Galliot, C., Hoballah, M., Kuhlemeier, C., and Stuurman, J. (2006). Genetics of flower size and nectar volume in Petunia pollination syndromes. Planta 225:203-212.

Gazzarrini, S., and McCourt, P. (2001). Genetic interactions between ABA, ethylene and sugar signaling pathways. Curr. Opin. Plant Biol. 4:387-391.

Gilad, Y., Pritchard, J., and Thornton, K. (2009). Characterizing natural variation using next-generation sequencing technologies. Trends Genet. 25:463-471.

Gillies, C.B. (1983). Ultrastructural studies of the association of homologous and non-homologous parts of chromosomes in the mid-prophase of meiosis in *Zea mays*. Maydica 28:265-287.

Gomez-Alvarez, V., Teal, T.K., and Schmidt, T.M. (2009). Systematic artifacts in metagenomes from complex microbial communities. ISME J. 3:1314-1317.

Gonzalez-Ibeas, D., Blanca, J., Roig, C., González-To, M., Picó, B., Truniger, V., Gómez, P., Deleu, W., Caño-Delgado, A., Arús, P., Nuez, F., Garcia-Mas, J., Puigdomènech, P., and Aranda, M., (2007). MELOGEN: an EST database for melon functional genomics. BMC Genomics 8:1-17.

González, A.M., Marcel, T.C., Kohutova, Z., Stam, P., Van der Linden, C.G., and Niks, R.E. (2010). Peroxidase Profiling Reveals Genetic Linkage between Peroxidase Gene Clusters and Basal Host and Non-Host Resistance to Rusts and Mildew in Barley. PLoS ONE 5:e10495.

Grattapaglia, D., Silva-Junior, O., Kirst, M., de Lima, B., Faria, D., and Pappas, G. (2011). High-throughput SNP genotyping in the highly heterozygous genome of Eucalyptus: assay success, polymorphism and transferability across species. BMC Plant Biol. 11:65.

Griffin, P., Robin, C., and Hoffmann, A. (2011). A next-generation sequencing method for overcoming the multiple gene copy problem in polyploid phylogenetics, applied to Poa grasses. BMC Biology 9:19.

Groenen, M.A.M., Cheng, H.H., Bumstead, N., Benkel, B.F., Briles, W.E., and Burke, T. (2000). A consensus linkage map of the chicken genome. Genome 10:137-147.

Gulyani, V., and Khurana, P. (2011). Identification and expression profiling of drought-regulated genes in mulberry (*Morus* sp.) by suppression subtractive hybridization of susceptible and tolerant cultivars. Tree Genetics and Genomes 7:725-738.

Gupta, S., Pandey-Rai, S., Srivastava, S., Naithani, S., Prasad, M., and Kumar, S. (2007). Construction of genetic linkage map of the medicinal and ornamental plant *Catharanthus roseus*. Genetics 86:259-268.

Han, T.H., de Jeu, M., van Eck, H., and Jacobsen, E. (2000). Genetic diversity of Chilean and Brazilian *Alstroemeria* species assessed by AFLP analysis. Heredity 84:564-569.

Hayashi, K., and Kawano, S. (2000). Molecular systematics of *Lilium* and allied genera (*Liliaceae*): phylogenetic relationships among *Lilium* and related genera based on the rbcL and matK gene sequence data. Plant Spec. Biol. 15:73-93.

He, B.Z., Holloway, A.K., Maerkl, S.J., and Kreitman, M. (2011). Does Positive Selection Drive Transcription Factor Binding Site Turnover? A Test with Drosophila Cis-Regulatory Modules. PLoS Genet 7:e1002053.

Heckenberger, M., Muminovic, J., Voort, J., Peleman, J., Bohn, M., and Melchinger, A.E. (2006). Identification of Essentially Derived Varieties Obtained from Biparental Crosses of Homozygous Lines. III. AFLP Data from Maize Inbreds and Comparison with SSR Data. Mol. Breed. 17:111-125.

Heled, J., and Drummond, A.J. (2010). Bayesian Inference of Species Trees from Multilocus Data. Mol. Biol. Evol. 27:570-580.

Hildebrand, D.F. (1989). Lipoxygenases. Physiol. Plant 76:249-253.

Hoeberichts, F.A., van Doorn, W.G., Vorst, O., Hall, R.D., and van Wordragen, M.F. (2007). Sucrose prevents up-regulation of senescence-associated genes in carnation petals. J. Exp. Bot. 58:2873-2885.

Huang, H., Lu, J., Ren, Z., Hunter, W., Dowd, S., and Dang, P. (2011). Mining and validating grape (*Vitis* L.) ESTs to develop EST-SSR markers for genotyping and mapping. Mol. Breed. 28:241-254.

Huang, X., and Madan, A. (1999). CAP3: A DNA Sequence Assembly Program. Genome Res. 9:868-877.

Hughes, C.E., Eastwood, R.J., and Bailey, C.D. (2006). From Famine to Feast? Selecting Nuclear DNA Sequence Loci for Plant Species-Level Phylogeny Reconstruction. Philos. Trans. R. Soc. London B. Biol. Sci. 361:211-225.

Huson, D.H., and Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies. Mol. Biol. Evol. 23:254-267.

Hyten, D., Cannon, S., Song, Q., Weeks, N., Fickus, E., Shoemaker, R., Specht, J., Farmer, A., May, G., and Cregan, P. (2010). High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11:38.

Iwaya-Inoue, M., and Takata, M. (2001). Trehalose plus Chloramphenicol Prolong the Vase Life of Tulip Flowers. HortScience 36:946-950.

Izawa, T., Takahashi, Y., and Yano, M. (2003). Comparative biology comes into bloom: genomic and genetic comparison of flowering pathways in rice and *Arabidopsis*. Curr. Opin. Plant Biol. 6:113-120.

Jaccoud, D., Peng, K., Feinstein, D., and Kilian, A. (2001). Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res. 29:e25.

Janiak, A., Kim, M., Van, K., and Lee, S.H. (2008). Application of degenerate oligonucleotide primed PCR (DOP-PCR) for SNP discovery in soybean. Euphytica 162:249-256.

Jiannis, R. (2006). Genotyping technologies for all. Drug Discovery Today: Technologies 3:115-122.

Joly, S., and Bruneau, A. (2006). Incorporating Allelic Variation for Reconstructing the Evolutionary History of Organisms from Multiple Genes: An Example from Rosa in North America. Syst. Biol. 55:623-636.

Jones, C.J., Edwards, K. J., Castaglione, S., Winfield, M. O., Sala, F., van de Wiel, C., Bredemeijer, G., Vosman, B., Matthes, M., Daly, A., Brettschneider, R., Bettini, P., Buiatti, M., Maestri, E., Malcevschi, A., Marmiroli, N., Aert, R., Volckaert, G., Rueda, J., Linacero, R., Vazquez, A., and Karp, A. (1997). Reproducibility testing of RAPD, AFLP and SSR markers in plants by a network of European laboratories. Mol. Breed. 3:381-390.

Jordan, B., Charest, A., Dowd, J.F., Blumenstiel, J.P., Yeh, R.-f., Osman, A., Housman, D.E., and Landers, J.E. (2002). Genome complexity reduction for SNP genotyping analysis. Proc. Natl Acad. Sci. USA 99:2942-2947.

Kamo, K., Blowers, A., Smith, F., Van Eck, J., and Lawson, R. (1995). Stable Transformation of Gladiolus Using Suspension Cells and Callus. J. Amer. Soc. Hort. Sci. 120:347-352.

Kamo, K., and Han, B.H. (2008). Biolistic-mediated Transformation of *Lilium* longiflorum cv. Nellie White. HortScience 43:1864-1869.

Karlov, G.I., L.I. Khrustaleva, K.B. Lim & J.M. Van Tuyl. (1999). The use of genomic *in situ* hybridization (GISH) to examine introgression and mechanism of 2n-pollen production in interspecific hybrids of lily. Genome 42:681-686.

Kawase, D., Hayashi, K., Takeuchi, Y., and Yumoto, T. (2010). Population genetic structure of *Lilium* japonicum and serpentine plant *L. japonicum* var. abeanum by using developed microsatellite markers. Plant Biosyst. 144:29-37.

Kentner, E.K., Arnold, M.L., and Wessler, S.R. (2003). Characterization of High-Copy-Number Retrotransposons From the Large Genomes of the Louisiana Iris Species and Their Use as Molecular Markers. Genetics 164:685-697.

Khan, N. (2009). A molecular cytogenetic study of intergenomic recombination and introgression of chromosomal segments in lilies (*Lilium*). PhD thesis, Wageningen University . The Netherlands. ISBN 78-90-8585-380-0,pp 121.

Khan, N., Barba-Gonzalez, R., Ramanna, M.S., Visser, R.G.F., and Van Tuyl, J.M. (2009). Construction of chromosomal recombination maps of three genomes of lilies (*Lilium*) based on GISH analysis. Genome 52:238-251.

Koes, R., Verweij, W., and Quattrocchio, F. (2005). Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. Trends Plant Sci. 10:236-242.

Konishi, T., Yano, Y., and Abe, K. (1992). Geographic distribution of alleles at the Ga2 locus for segregation distortion in barley. Theor. Appl. Genet. 85:419-422.

Kozarewa, I., Ning, Z., Quail, M.A., Sanders, M.J., Berriman, M., and Turner, D.J. (2009). Amplification-free Illumina sequencing-library preparation facilitates improved mapping and assembly of (G+C)-biased genomes. Nat. Meth. 6:291-295.

Krens, F., and Van Tuyl, J.M. (2011). Plant breeding in bulbous ornamentals: Adding wit to chance. Acta Hort. 886:329-342.

Ku, H.-M., Vision, T., Liu, J., and Tanksley, S.D. (2000). Comparing sequenced segments of the tomato and Arabidopsis genomes: Large-scale duplication followed by selective gene loss creates a network of synteny. Proc. Natl. Acad. Sci. USA 97:9121-9126.

Kumar, S., and Blaxter, M. (2010). Comparing de novo assemblers for 454 transcriptome data. BMC Genomics 11:571.

Kumaran, M.K., Ye, D., Yang, W.-C., Griffith, M.E., Chaudhury, A.M., and Sundaresan, V. (1999). Molecular cloning of ABNORMAL FLORAL ORGANS: a gene required for flower development in *Arabidopsis*. Sex. Plant Reprod. 12:118-122.

Lagercrantz, U., and Lydiate, D.J. (1996). Comparative Genome Mapping in *Brassica*. Genetics 144:1903-1910.

Lan, T.H., DelMonte, T.A. Reischmann, K.P. Hyman, J. Kowalski, S.P. McFerson, J. Kresovich, S. and Paterson A.H. (2000). An EST-enriched Comparative Map of *Brassica oleracea* and *Arabidopsis thaliana*. Genome Res. 10:776-788.

Lange, K., and Boehnke, M. (1982). How many polymorphic genes will it take to span the human genomic? Am. J. Hum. Genet. 34:842-845.

León, P., and Sheen, J. (2003). Sugar and hormone connections. Trends Plant Sci. 8:110-116.

Lepoittevin, C., Frigerio, J.-M., Garnier-Géré, P., Salin, F., Cervera, M.T., Vornam, B., Harvengt, L., and Plomion, C.(2010). *In Vitro* vs *In Silico* Detected SNPs for the Development of a Genotyping Array: What Can We Learn from a Non-Model Species? PLoS ONE 5:e11034.

Leslie, A.C. (1982). The International Lily Register and the 24 Supplements. Roy. Hort. Soc. London. http://www.lilyregister.com.

Li, M., Wunder, J., Bissoli, G., Scarponi, E., Gazzani, S., Barbaro, E., Saedler, H., and Varotto C. (2008). Development of COS genes as universally amplifiable markers for phylogenetic reconstructions of closely related plant species. Cladistics 24:727-745.

Li, W., and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.

Lim, K.B., Wennekes, J., De Jong, J.H., Jacobsen, E., and Van Tuyl, J.M. (2001). Karyotype analysis of *Lilium longiflorum* and *Lilium rubellum* by chromosome banding and fluorescence *in situ* hybridisation. Genome 44:911-918.

Lim, K.B. (2000). Introgression breeding through interspecific polyploidisation in lily: a molecular cytogenetic study: Wageningen University. ISBN 9789058083111, pp116.

Lim, K.B., and Van Tuyl, J.M. (2006). *Lilium* hybrids. In: Flower breeding and genetics: issues, challenges and opportunities for the 21st century--Anderson, N.O., ed.: Springer, Dordrecht, the Netherlands 517–537.

Lin, H., Ouyang, S., Egan, A., Nobuta, K., Haas, B., Zhu, W., Gu, X., Silva, J., Meyers, B., and Buell, C.R. (2008). Characterization of paralogous protein families in rice. BMC Plant Biol. 8:18.

Liu, F., W. Xu, Q. Wei, Z. Zhang, Z. Xing, L. Tan, C. Di, D. Yao, C. Wang, Y. Tan, H. Yan, Y. Ling, C. Sun, Y. Xue, and Su Z. (2010). Gene Expression Profiles Deciphering Rice Phenotypic Variation between Nipponbare (Japonica) and 93-11 (Indica) during Oxidative Stress. PLoS ONE 5:e8632.

Liu, L., Pearl, D., Brumfield, R., and Edwards, S. (2008). Estimating species trees using multiple-allele DNA sequence data. Evolution 62:2080-2091.

Livingstone, K.D., Lackney, V.K., Blauth, J.R., van Wijk, R., and Jahn, M.K. (1999). Genome Mapping in Capsicum and the Evolution of Genome Structure in the *Solanaceae*. Genetics 152:1183-1202.

Lu, C., and Bridgen, M.P. (1996). Effects of genotype, culture medium and embryo developmental stage on the in vitro responses from ovule cultures of interspecific hybrids of *Alstroemeria*. Plant Sci. 116:205-212.

Lu, G., Zhang, X., Zou, Y., Zou, Q., Xiang, X., and Cao, J. (2007). Effect of radiation on regeneration of Chinese narcissus and analysis of genetic variation with AFLP and RAPD markers. Plant Cell Tissue Organ Cult. 88:319-327.

Lynch, M., and Conery, J.S. (2000). The Evolutionary Fate and Consequences of Duplicate Genes. Science 290:1151-1155.

Ma, J., Morrow, D., Fernandes, J., and Walbot, V. (2006). Comparative profiling of the sense and antisense transcriptome of maize lines. Genome Biol. 7:1-18.

Mace, E., Xia, L., Jordan, D., Halloran, K., Parh, D., Huttner, E., Wenzl, P. and Kilian, A., (2008). DArT markers: diversity analyses and mapping in *Sorghum bicolor*. BMC Genomics 9:26.

Maddison, W.P., and Maddison, D.R. (2011). Mesquite: a modular system for evolutionary analysis. Version 2.75. http://mesquiteproject.org.

Mammadov, J., Chen, W., Ren, R., Pai, R., Marchione, W., Yalçin, F., Witsenboer, H., Greene, T., Thompson, S., and Kumpatla, S. (2010). Development of highly polymorphic SNP markers from the complexity reduced portion of maize *Zea mays* L. genome for use in marker-assisted breeding. Theor. Appl. Genet. 121:577-588.

Mantovani, P., Maccaferri, M., Sanguineti, M., Tuberosa, R., Catizone, I., Wenzl, P., Thomson, B., Carling, J., Huttner, E., DeAmbrogio, E., and Kilian, A. (2008) An integrated DArT-SSR linkage map of durum wheat. Mol. Breed. 22: 629-648.

Marasek, A., Hasterok, R., Wiejacha, K., and Orlikowska, T. (2004). Determination by GISH and FISH of hybrid status in *Lilium*. Hereditas 140:1-7.

Martin, B., Nienhuis, J., King, G., and Schaefer, A. (1989). Restriction fragment length polymorphisms associated with water use efficiency in tomato. Science 243:1725-1728.

Martin, J.A., and Wang, Z. (2011). Next-generation transcriptome assembly. Nat. Rev. Genet. 12:671-682.

Martin, N.H., Bouck, A.C., and Arnold, M.L. (2007). The Genetic Architecture of Reproductive Isolation in Louisiana Irises: Flowering Phenology. Genetics 175:1803-1812.

Martin, N.H., Y. Sapir, and Arnold, M.L. (2008). The genetic architecture of reproductive isolation in Louisiana irises: pollination syndromes and pollinator preferences. Evolution 62: 740-752.

Mayak, S., and Dilley, D.R. (1976). Effect of sucrose on response of cut carnation to kinetin, ethylene, and abscisic acid. J. Amer. Sot. Hort. Sci. 101:583-585.

McCouch, S.R. (2001). Genomics and Synteny. Plant Physiol. 125:152-155.

McIntosh, K.B., and Bonham-Smith, P.C. (2001). Establishment of *Arabidopsis thaliana* ribosomal protein RPL23A-1 as a functional homologue of *Saccharomyces cerevisiae* ribosomal protein L25. Plant Mol.Biol. 46:673-682.

McRae, E.A (1998a). Lilies: a guide for growers and collectors. Portland, Oregon: Timber press. ISBN 10:0881924105, pp 392.

McRae, E.A. (1998b). Lily species. In: Lilies. Timber Press. Portland, Oregon. ISBN 10:0881924105,105-204.

McRae, E.A. (1990). American lily hybridising- an historical review. In: Hayward AF (ed) Lilies and related plants, suppl. Roy Hort. Soc-Lily. 29-40.

Messmer, M.M., Keller, M., Zanetti, S., and Keller, B. (1999). Genetic linkage map of a wheat×spelt cross. Theor. Appl. Genet. 98:1163-1170.

Miller, M.A., Pfeiffer, W., and Schwartz, T. (2010). Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: Gateway Computing Environments Workshop (GCE), 2010. 1-8.

Miller, M.R., Dunham, J.P., Amores, A., Cresko, W.A., and Johnson, E.A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res. 17:240-248.

Milne, I., Lindner, D., Bayer, M., Husmeier, D., McGuire, G., Marshall, D.F., and Wright, F. (2009). TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. Bioinformatics 25:126-127.

Mizuochi, H., Marasek, A., and Okazaki, K. (2007). Molecular cloning of *Tulipa fosteriana* rDNA and subsequent FISH analysis yields cytogenetic organization of 5S rDNA and 45S rDNA in *T. gesneriana* and *T. fosteriana*. Euphytica 155:235-248.

Moore, S., Payton, P., Wright, M., Tanksley, S., and Giovannoni, J. (2005). Utilization of tomato microarrays for comparative gene expression analysis in the *Solanaceae*. J. Exp. Bot. 56:2885-2895.

Morgante, M., and Olivieri, A.M. (1993). PCR-amplified microsatellites as markers in plant genetics. Plant J. 3:175-182.

Müller, R., Stummann, B.M., Andersen, A.S., and Serek, M. (1999). Involvement of ABA in postharvest life of miniature potted roses. Plant Growth Regul. 29:143-150.

Muratović, E., Hidalgo, O., Garnatje, T., and Siljak-Yakovlev, S. (2010). Molecular phylogeny and genome size in European lilies (Genus *Lilium*, *Liliaceae*). Advanced Science Letters. 3:180-189.

Myles, S., Boyko, A.R., Owens, C.L., Brown, P.J., Grassi, F., Aradhya, M.K., Prins, B., Reynolds, A., Chia, J.-M., Ware, D., Bustamante, C.D. and Buckler, E.S. (2011). Genetic structure and domestication history of the grape. Proc. Natl Acad. Sci. USA 108:3530-3535.

Myles, S., Peiffer, J., Brown, P.J., Ersoz, E.S., Zhang, Z., Costich, D.E., and Buckler, E., (2009). Association mapping: Critical considerations shift from genotyping to experimental design. Plant Cell 21:2194-2202.

Nadeau, J.A., Zhang, X.S., Nair, H., and O'Neill, S.D. (1993). Temporal and spatial regulation of 1-aminocyclopropane-1-carboxylate oxidase in the pollination-induced senescence of orchid flowers. Plant Physiol. 103:31-39.

Nakano, M., Nakatsuka, A., Nakayama, M., Koshioka, M., and Yamagishi, M. (2005). Mapping of quantitative trait loci for carotenoid pigmentation in flower tepals of Asiatic hybrid lily. Sci Horti-Amsterdam 104:57-64.

Narina, S.S., Buyyarapu, R., Kottapalli, K.R., Sartie, A.M., Ali, M.I., Robert, A., Hodeba, M.J.D., Sayre, B.L., and Scheffler, B.E. (2011). Generation and analysis of expressed sequence tags (ESTs) for marker development in yam (*Dioscorea alata* L.). BMC Genomics 12:100.

Nei, M. (2005). Selectionism and Neutralism in Molecular Evolution. Mol. Biol. Evol. 22:2318-2342.

Niedringhaus, T., Milanova, D., Kerby, M., Snyder, M., and Barron, A. (2011). Landscape of next-generation sequencing technologies. Anal. Chem. 83:4327-4341.

Nishikawa, T., Okazaki, K., Arakawa, K., and Nagamine, T. (2001). Phylogenetic Analysis of Section Sinomartagon in Genus *Lilium* Using Sequences of the Internal Transcribed Spacer Region in Nuclear Ribosomal DNA. Breed. Sci. 51:39-46.

Nishikawa, T., Okazaki, K., Uchino, T., Arakawa, K., and Nagamine, T. (1999). A Molecular Phylogeny of *Lilium* in the Internal Transcribed Spacer Region of Nuclear Ribosomal DNA. J. Mol. Evol. 49:238-249.

Nishiyama, T., Fujita, T., Shin-I, T., Seki, M., Nishide, H., Uchiyama, I., Kamiya, A., Carninci, P., Hayashizaki, Y., Shinozaki, K., Kohara, Y., and Hasebe, M. (2003). Comparative genomics of Physcomitrella patens gametophytic transcriptome and *Arabidopsis thaliana*: Implication for land plant evolution. Proc. Natl Acad. Sci. USA 100:8007-8012.

Novaes, E., Drost, D., Farmerie, W., Pappas, G., Grattapaglia, D., Sederoff, R., and Kirst, M. (2008). High-throughput gene and SNP discovery in *Eucalyptus grandis*, an uncharacterized genome. BMC Genomics 9:312.

Nowak, J., and Mynett, K. (1985). The effect of sucrose, silver thiosulphate and 8-hydroxyquinoline citrate on the quality of *Lilium* inflorescences cut at the bud stage and stored at low temperature. Sci. Hortic. 25:299-302.

Nozawa, M., Suzuki, Y., and Nei, M. (2009). Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc. Natl Acad. Sci. USA 106:6700-6705.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., and Nieves-Aldrey, J. (2004). Bayesian Phylogenetic Analysis of Combined Data. Syst. Biol. 53:47-67.

Olmstead, R., and Palmer, J. (1992). A chloroplast DNA phylogeny of the Solanaceae: Subfamilial relationships and character evolution. Ann. Missouri. Bot. Gard. 79:346 - 360.

Ozdemir Ozgenturk, N., Oruç, F., Sezerman, U., Kuçukural, A., Vural Korkut, S., Toksoz, F., and Un, C. (2010). Generation and Analysis of Expressed Sequence Tags from *Olea europaea* L. Compar. Funct. Genom. 2010:9.

Page, R.D. (1998). GeneTree: comparing gene and species phylogenies using reconciled trees. Bioinformatics 14:819-820.

Palmieri, N., and Schlötterer, C. (2009). Mapping Accuracy of Short Reads from Massively Parallel Sequencing and the Implications for Quantitative Expression Profiling. PLoS ONE 4:e6323.

Panavas, T., Walker, E.L., and Rubinstein, B. (1998). Possible involvement of abscisic acid in senescence of daylily petals. J. Exp. Bot. 49:1987-1997.

Papanicolaou, A., Stierli, R., ffrench-Constant, R., and Heckel, D. (2009). Next generation transcriptomes for next generation genomes using est2assembly. BMC Bioinformatics 10:447.

Parchman, T., Geist, K., Grahnen, J., Benkman, C., and Buerkle, C.A. (2010). Transcriptome sequencing in an ecologically important tree species: assembly, annotation, and marker discovery. BMC Genomics 11:180.

Park, M., Jo, S., Kwon, J.-K., Park, J., Ahn, J.H., Kim, S., Lee, Y.-H., Yang, T.-J., Hur, C.-G., Kang, B.-C., Kim, B.-D., and Choi, D., (2011). Comparative analysis of pepper and tomato reveals euchromatin expansion of pepper genome caused by differential accumulation of Ty3/Gypsy-like elements. BMC Genomics 12:85.

Paszkiewicz, K., and Studholme, D. (2010). *De novo* assembly of short sequence reads. Brief. Bioinform. 11:457-472.

Pavord, A. (1999). The tulip: Bloomsbury USA. ISBN-10: 1582341303, pp 296.

Pavy, N., Pelgas, B., Beauseigle, S., Blais, S., Gagnon, F., Gosselin, I., Lamothe, M., Isabel, N., and Bousquet, J. (2008). Enhancing genetic mapping of complex genomes through the design of highly-multiplexed SNP arrays: application to the large and unsequenced genomes of white spruce and black spruce. BMC Genomics 9:21.

Petersen, L., Bollback, J.P., Dimmic, M., Hubisz, M., and Nielsen, R. (2007). Genes under positive selection in Escherichia coli. Genome Res. 17:1336-1343.

Phatak, S.C., Wittwer, S.H., Honma, S., and Bukovac, M.J. (1966). Gibberellin-induced Anther and Pollen Development in a Stamen-less Tomato Mutant. Nature 209:635-636.

Pillen K, Steinrucken G, Wricke G, Herrmann RG, and Jung, C. (1993). A extended linkage map of sugar beet (*Beta vulgaris* L.) including nine putative lethal genes and restorer gene X. Plant Breed. 111:265-272.

Popowich, E.A., Firsov, A.P., Mitiouchkina, T.Y., Filipenya, V.L., Dolgov, S.V., and Reshetnikov, V.N. (2007). *Agrobacterium*-mediated transformation of *Hyacinthus orientalis* with thaumatin II gene to control fungal diseases. Plant Cell Tissue Organ Cult. 90:237-244.

Pourtau, N., Marès, M., Purdy, S., Quentin, N., Ruël, A., and Wingler, A. (2004). Interactions of abscisic acid and sugar signalling in the regulation of leaf senescence. Planta 219:765-772.

Qiu, L., Yang, C., Tian, B., Yang, J.-B., and Liu, A. (2010). Exploiting EST databases for the development and characterization of EST-SSR markers in castor bean (*Ricinus communis* L.). BMC Plant Biol. 10:278.

Rahman, M.u., Asif, M., Shaheen, T., Tabbasam, N., Zafar, Y., Paterson, A.H., and Lichtfouse, E. (2011). Marker-Assisted Breeding in Higher Plants. In: Alternative Farming Systems, Biotechnology, Drought Stress and Ecological Fertilisation--Lichtfouse, E., ed.: Springer Netherlands 39-76.

Ramanna, M.S., and Jacobsen, E. (2003). Relevance of sexual polyploidization for crop improvement - A review. Euphytica 133:3-18.

Ramon, M., Rolland, F., and Sheen, J. (2008). Sugar Sensing and Signaling. The *Arabidopsis* Book:e0117.

Ranjan, P., Bhat, K.V., Misra, R.L., Singh, S.K., and Ranjan, J.K. (2010). Genetic relationships of gladiolus cultivars inferred from fluorescence based AFLP markers. Sci. Hortic. 123:562-567.

Rešetnik, I., Liber, Z., Satovic, Z., Cigić, P., and Nikolić, T. (2007). Molecular phylogeny and systematics of the *Lilium carniolicum* group (*Liliaceae*) based on nuclear ITS sequences. Plant Syst. Evol. 265:45-58.

Rhee, H.K., Lim, J.H., Kim, Y.J., and Van Tuyl, J.M. (2005). Improvement of breeding efficiency for interspecific hybridization of lilies in Korea. Acta Hort. 673:107 -112

Ridgeway, J. (2004). It's all for sale: the control of global resources. Durham: Duke University Press. ISBN 0-8223-3374-0, pp 250.

Rivarola, M., Foster, J.T., Chan, A.P., Williams, A.L., Rice, D.W., Liu, X., Melake-Berhan, A., Creasy, H.H., Puiu, D., Rosovitz, M.J., Khouri, H.M., Beckstrom-Sternberg, S.M., Allan, G.J., Keim, P., Ravel, J., and Rabinowicz, P.D. (2011). Castor Bean Organelle genome sequencing and worldwide genetic diversity analysis. PLoS ONE 6(7): e21743.

Rodionov, A.V., Lukina, N.A., Galkina, S.A., Solovei, I., and Saccone, S. (2002). Crossing over in chicken oogenesis: Cytological and chiasma-based genetic maps of the chicken lampbrush chromosome 1. J. Hered. 93:125-129.

Rota, M., and Sorrells, M. (2004). Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. Funct. Integr. Genomics 4:34-46.

Roth, C., and Liberles, D. (2006). A systematic search for positive selection in higher plants (Embryophytes). BMC Plant Biol. 6:12.

Rubio-Moraga, A., Castillo-Lopez, R., Gomez-Gomez, L., and Ahrazem, O. (2009). Saffron is a monomorphic species as revealed by RAPD, ISSR and microsatellite analyses. BMC Research Notes 2:189.

Sakai, H., Ikawa, H., Tanaka, T., Numa, H., Minami, H., Fujisawa, M., Shibata, M., Kurita,. Kikuta, K., A, Hamada, M., Kanamori, H., Namiki, N., Wu, J., Itoh, T., Matsumoto, T., and Sasaki, T. (2011). Distinct evolutionary patterns of *Oryza glaberrima* deciphered by genome sequencing and comparative analysis. Plant. J. 66:796-805.

Salse, J., Piégu, B., Cooke, R., and Delseny, M. (2002). Synteny between *Arabidopsis thaliana* and rice at the genome level: a tool to identify conservation in the ongoing rice genome sequencing project. Nucleic Acids Res. 30:2316-2328.

Sanderson, M., and McMahon, M. (2007). Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 7:S3.

Sato, T., and Miyoshi, K. (2007). Restoration of intact anthers in a thermosensitive, antherless, malesterile cultivar of Asiatic hybrid lily in response to high temperature. J. Hortic. Sci. Biotech. 82:791-797.

Schaal, B.A., Hayworth, D.A., Olsen, K.M., Rauscher, J.T., and Smith, W.A. (1998). Phylogeographic studies in plants: problems and prospects. Mol. Ecol. 7:465-474.

Schneeberger, K., and Weigel, D. (2011). Fast-forward genetics enabled by new sequencing technologies. Trends Plant Sci. 16:282-288.

Schroeder, J.I., Kwak, J.M., and Allen, G.J. (2001). Guard cell abscisic acid signalling and engineering drought hardiness in plants. Nature 410:327-330.

Scientists, G.K.C.o. (2009). Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species. J. Hered. 100:659-674.

Shahin, A., Arens, P., Van Heusden, A.W., Van der Linden, G., Van Kaauwen, M., Khan, N., Schouten, H.J., Van De Weg, W.E., Visser, R.G.F., and Van Tuyl, J.M. (2011). Genetic mapping in *Lilium*: mapping of major genes and quantitative trait loci for several ornamental traits and disease resistances. Plant Breed. 130:372-382.

Shahin, A., Arens, P., Van Heusden, S., and Van Tuyl, J.M. (2009). Conversion of molecular markers linked to *Fusarium* and Virus resistance in Asiatic lily hybrids. Acta Hort. 836:131-136.

Shahin, A., Van Gurp, T., Peters, S.A., Visser, R.G.F., Van Tuyl, J.M., and Arens, P. (2012). SNP markers retrieval for a non-model species: a practical approach. BMC Research Notes 5:79.

Shen, Y.Y., Wang, X.-F., Wu, F.-Q,. Du, S.-Y., Cao, Z., Shang, Y., Wang, X.-L., Peng, C.-C., Yu, X.-C., Zhu, S.-Y., Fan, R.-C., Xu, Y.-H., and Zhang, D.-P. (2006). The Mg-chelatase H subunit is an abscisic acid receptor. Nature 443:823-826.

Shimizu, M. (1987). The lilies of Japan (In Japanese). Seibundo Shinkosha. Tokyo, pp.148–165.

Sibov S.T., Souza, C.L.J.R., Garcia, A.A.F., Garcia, A.F., Silva, A.R., Mangolin, C.A., Benchimol, L.L., and Souza, A. (2003). Molecular mapping in tropical maize (*Zea mays* L.) using microsatellite

markers. 1. Map construction and localization of loci showing distorted segregation. Hereditas 139:96-106.

Simón, V.I., Picó, F.X., and Arroyo, J. (2010). New microsatellite loci for *Narcissus papyraceus* (*Amarillydaceae*) and cross-amplification in other congeneric species. Am. J. Bot. 97:e10-e13.

Singhal, D., Gupta, P., Sharma, P., Kashyap, N., Anand, S., and Sharma, H. (2011). *In-silico* single nucleotide polymorphisms (SNP) mining of *Sorghum bicolor* genome. AFR. J. Biotechnol. 10:580 -583

Small, R.L., Cronn, R.C., and Wendel, J.F. (2004). Use of nuclear genes for phylogeny reconstruction in plants. Aust. Syst. Bot. 17:145-170.

Smith, J.S.C., Chin, E.C.L., Shu, H., Smith, O.S., Wall, S.J., Senior, M.L., Mitchell, S.E., Kresovich, S., aand Ziegle, J. (1997). An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparisons with data from RFLPS and pedigree. Theor. Appl. Genet. 95:163-173.

Smulders, M.J.M., Vukosavljev, M., Shahin, A., van de Weg, W., and Arens, P. (submitted). High throughput marker development and application in horticultural crops. Acta Hort.

Song, C., Bang, C., Chung, S., Kim, Y., Lee, J. and Lee, D. . (1996). Effects of postharvest pretreatments and preservative solutions on vase life and flower quality of Asiatic hybrid lily. Acta Hort. 414:277-286.

Stack, S.M., Anderson, L.K., and Sherman, J.D. (1989). Chiasmata and recombination nodules in *Lilium* longiflorum. Genome 32:486.

Stamatakis, A. (2006). Phylogenetic models of rate heterogeneity: a high performance computing perspective. In: Parallel and Distributed Processing Symposium, 2006. IPDPS 2006. 20th International. 8 pp.

Stamatakis, A., Hoover, P., and Rougemont, J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web Servers. Syst. Biol. 57:758-771.

Steele, P.R., and Pires, J.C. (2011). Biodiversity assessment: State-of-the-art techniques in phylogenomics and species identification. Am. J. Bot. 98:415-425.

Sterck, L., Rombauts, S., Vandepoele, K., Rouzé, P., and Van de Peer, Y. (2007). How many genes are there in plants (... and why are they there)? Curr. Opin. Plant Biol. 10:199-203.

Stewart, R.N. (1947). The Morphology of Somatic Chromosomes in *Lilium*. Am. J. Bot. 34:9-26.

Straathof, T.h.P., Van Tuyl, J.M., Dekker, B., Van Winden, M.J.M., and Sandbrink, J.M. (1996). Genetic analysis of inheritance of partial resistance to *Fusarium oxysporum* in Asiatic hybrids of lily using RAPD markers. Acta Hort. 414:209-218.

Straathof, T.P., and Loffler, H.J.M. (1994). Resistance to *Fusarium oxysporum* at Different Developmental Stages of Asiatic Hybrid Lilies. J. Amer. Soc. Hort. Sci. 119:1068-1072.

Straub, S.C.K., Parks, M., Weitemier, K., Fishbein, M., Cronn, R.C., and Liston, A. (2012). Navigating the tip of the genomic iceberg: Next-generation sequencing for plant systematics. Am. J. Bot. 99:349 –364.

Swart, A. (1981). Quality of *Lilium* "Enchantment" flowers as influenced by season and silver thiosulphate. Acta Hort. 113:45-50.

Swofford, D.L. (2003). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts, USA.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28:2731–2739.

Tang, H., Lyons, E., Pedersen, B., Schnable, J., Paterson, A., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics 12:102.

Tang, J., Leunissen, J., Voorrips, R., Van der Linden, C.G., and Vosman, B. (2008). HaploSNPer: a web-based allele and SNP detection tool. BMC Genetics 9:23.

Tang, J., Vosman, B., Voorrips, R., Van der Linden, C.G., and Leunissen, J. (2006). QualitySNP: a pipeline for detecting single nucleotide polymorphisms and insertions/deletions in EST data from diploid and polyploid species. BMC Bioinformatics 7:438.

Tang, S., Okashah, R., Cordonnier-Pratt, M.-M., Pratt, L., Ed Johnson, V., Taylor, C., Arnold, M., and Knapp, S. (2009). EST and EST-SSR marker resources for Iris. BMC Plant Biol. 9:72.

Tang, X., Gomes, A.M.T.R., Bhatia, A., and Woodson, W.R. (1994). Pistil-specific and ethylene-regulated expression of 1-aminocyclopropane-1-carboxylate oxidase genes in petunia flowers. Plant Cell 6:1227-1239.

Tang, X., and Woodson, W.R. (1996). Temporal and spatial expression of 1-aminocyclopropane-1-carboxylate oxidase mRNA following pollination of immature and mature petunia flowers. Plant Physiol. 112:503-511.

Tanhuanpää, P. (2004). Identification and mapping of resistance gene analogs and a white rust resistance locus in *Brassica rapa* ssp. oleifera. Theor. Appl. Genet. 108:1039-1046.

Tanksley, S.D., Ganal, M.W., Prince, J.P., De Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B., Messeguer, R., Miller, J.C., Miller, L., Paterson, A.H., Pineda, O., Roder, M.S., Wing, R.A., Wu, W., and Young, N.D. (1992). High density molecular linkage maps of the tomato and potato genomes. Genetics 132:1141-1160.

Ten Have, A., and Woltering, E.J. (1997). Ethylene biosynthetic genes are differentially expressed during carnation (*Dianthus caryophyllus* L.) flower senescence. Plant Mol. Biol. 34:89-97.

The Angiosperm Phylogeny Group. (2003). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II. Bot. J. Linnean Soc. 141:399-436.

The Angiosperm Phylogeny Group. (2009). An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III. Bot. J. Linnean Soc. 161:105-121.

Thiel, Michalek, Varshney, and Graner. (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.) Theor. Appl. Genet. 106:411-422.

Thorup, T.A., Tanyolac, B., Livingstone, K.D., Popovsky, S., Paran, I., and Jahn, M. (2000). Candidate gene analysis of organ pigmentation loci in the *Solanaceae*. Proc. Natl. Acad. Sci. USA. 97:11192-11197.

Tillett, R.L., Ergül, A., Albion, R.L., Schlauch, K.A., Cramer, G.R., and Cushman, J.C. (2011). Identification of tissue-specific, abiotic stress-responsive gene expression patterns in wine grape (*Vitis vinifera* L.) based on curation and mining of large-scale EST data sets. BMC Plant Biol. 11:86.

Treangen, T.J., and Salzberg, S.L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat. Rev. Genet. 13:146-146.

Tripathi, S.K., and Tuteja, N. (2007). Integrated signaling in flower senescence: An overview. Plant Signaling and Behavior 2:437-445.

Trivellini, A., Ferrante, A., Vernieri, P., Mensuali-Sodi, A., and Serra, G. (2011a). Effects of Promoters and Inhibitors of Ethylene and ABA on Flower Senescence of *Hibiscus rosa-sinensis* L. J. Plant Growth Regul. 30:175-184.

Trivellini, A., Ferrante, A., Vernieri, P., and Serra, G. (2011b). Effects of abscisic acid on ethylene biosynthesis and perception in *Hibiscus rosa-sinensis* L. flower development. J. Exp. Bot. 62:5437-5452.

Twyman, R.M., and Primrose, S.B. (2003). Techniques patents for SNP genotyping. Pharmacogenomics 4:67-79.

Van Creij, M.G.M., Kerckhoffs, D.M.F.J., and Van Tuyl, J.M. (1997). Interspecific crosses in the genus *Tulipa* L.: identification of pre-fertilization barriers. Sex. Plant Reprod. 10:116-123.

Van der Linden, C.G., Smulders, M.J.M., and Vosman, B.J. (2005). Motif-directed profiling: a glance at molecular evolution. In: Plant species-level systematics: new perspectives on pattern & process Liechtenstein: ARG Gantner Verlag. 291-303.

Van der Linden, C.G., Wouters, D.C.A.E., Mihalka, V., Kochieva, E.Z., Smulders, M.J.M., and Vosman, B. (2004). Efficient targeting of plant disease resistance loci using NBS profiling. Theor. Appl. Genet. 109:384-393.

Van der Meulen-Muisers, J.J.M. (2000). Genetic and physiological aspects of postharvest flower longevity in Asiatic hybrid lilies (*Lilium* L.): Wageningen Universiteit. ISBN 90-5808-276-8. pp119.

Van der Meulen-Muisers, J.J.M., Van Oeveren, J.C., Jansen, J., and Van Tuyl, J.M. (1999). Genetic analysis of postharvest flower longevity in Asiatic hybrid lilies. Euphytica 107:149-157.

Van der Meulen-Muisers, J.J.M., Van Oeveren, J.C., Van der Plas, L.H.W., and Van Tuyl, J.M. (2001). Postharvest flower development in Asiatic hybrid lilies as related to tepal carbohydrate status. Postharvest Biol. Technol. 21:201-211.

Van der Meulen-Muisers, J.J.M., Van Oeveren, J.C., and Van Tuyl, J.M. (1998). Genotypic Variation in Postharvest Flower Longevity of Asiatic Hybrid Lilies. J. Amer. Soc. Hort. Sci. 123:283-287.

Van der Meulen-Muisers, J.J.M., Van Oeveren, J.C., Meijkamp, B.B. and Derks, F.H.M. (1995). Effect of floral bud reduction on flower longevity in Asiatic hybrid lilies. Acta Hort. 405:46-57.

Van der Meulen, J.J.M., Van Oeveren, J.C., Sandbrink, J.M., and VanTuyl, J.M. (1996). Molecular markers as a tool for breeding for flower longevity in Asiatic hybrid lilies. Acta Hort. 420:68-71.

Van der Meulen, J.J.M., Van Oeveren, J.C., and Tuyl, J.M.V. (1997). Breeding as a tool for improving postharvest quality characters of lily and tulip flowers. Acta Hort. 430:569-575.

Van Doorn, W.G. (2001). Role of soluble carbohydrates in flower senescence: a survey. Acta Hort. 543:179-183.

Van Doorn, W.G. (2004). Is petal senescence due to sugar starvation? Plant Physiol. 134:35-42.

Van Doorn, W.G. (2011). The postharvest quality of cut lily flowers and potted lily plants. Acta Hort. 900:255-264.

Van Doorn, W.G., and Stead, A.D. (1994). The physiology of petal senescence which is not initiated by ethylene. In: scott RJ, Stead AD, eds. Molecular and Cellular Aspects of Plant Reproduction. Cambridge University Press: 239-254.

Van Doorn, W.G., and Woltering, E.J. (2008). Physiology and molecular biology of petal senescence. J. Exp. Bot. 59:453-480.

Van Harten, A. (2002). Mutation breeding of vegetatively propagated ornamentals. In: Breeding for ornamentals: classical and molecular approaches-Vainstein, A., ed.: Dordrecht; London: Kluwer Academic 105-127.

Van Heusden, A.W., Jongerius, M.C., Van Tuyl, J.M., Straathof, T.P., and Mes, J.J. (2002). Molecular assisted breeding for disease resistance in lily. Acta Hort. 572:131-138.

Van Ooijen, J. (2006). JoinMap ® 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma B.V., Wageningen, Netherlands.

Van Orsouw, N.J., Hogers, R.C.J., Janssen, A., Yalcin, F., Snoeijers, S., Verstege, E., Schneiders, H., Van der Poel, H., Van Oeveren, J., Verstegen, H., and Van Eijk, M.J.T. (2007). Complexity Reduction of Polymorphic Sequences (CRoPS™): A Novel Approach for Large-Scale Polymorphism Discovery in Complex Genomes. PLoS ONE 2:e1172.

Van Raamsdonk, L., W,D. (1992). Biosystematic studies in *Tulipa* subgenus Eriostemones. Plant Syst. Evol. 179:127-141.

Van Raamsdonk, L.W.D., and De Vries, T. (1992). Biosystematic studies in*Tulipa* sect. *Eriostemones* (*Liliaceae*) Plant Syst. Evol. 179:27-41.

Van Raamsdonk, L.W.D., and De Vries, T. (1995). Species Relationships and Taxonomy in Subg *Tulipa* (*Liliaceae*). Plant Syst. Evol. 195:13-44.

Van Raamsdonk, L.W.D., Eikelboom, W., De Vries, T., and Straathof, Th.P. (1997). The systematics of the genus *Tulipa* L Acta Hort. 430:821-828.

Van Tuyl, J.M., Arens, P., and Marasek-Ciolakowska, A., (in press). Breeding and Genetics of ornamental geophytes. In: Ornamental Geophytes: From Basic Science to Sustainable Horticultural Production--Kamenetsky, R., and Okubo, H., eds. Taylor & Francis.

Van Tuyl, J.M., Maas, I.W.G.M., and Lim, K.B. (2002). Introgression in interspecific hybrids of lily. Acta Hort. 570:213-218.

Van Tuyl, J.M., and Van Creij, M.G.M. (2006). *Tulipa gesneriana* and *T.* hybrids. In: Flower breeding and genetics: issues, challenges and opportunities for the 21st century--Anderson, N.O., ed.: Springer, Dordrecht, the Netherlands 623–641.

Van Tuyl, J.M., Van Diën, M.P., Van Creij, M.G.M., Van Kleinwee, T.C.M., Franken, J., and Bino, R.J. (1991). Application of in vitro pollination, ovary culture, ovule culture and embryo rescue for overcoming incongruity barriers in interspecific *Lilium* crosses. Plant Sci. 74:115-126.

Van Tuyl, J.M., Van Dijken, A., Chi, H.S., Lim, K.B., Villemoes, S., and Van Kronenburg, B.C.E. (2000). Breakthroughs in interspecific hybridization of lily. Acta Hort. 508:83-88.

Van Tuyl, J.M., and Lim, K.B. (2003). Interspecific hybridization and polyploidization as tools in ornamental plant breeding. Acta Hort. 612:13-22.

Varshney, R., Grosse, I., Hähnel, U., Siefken, R., Prasad, M., Stein, N., Langridge, P., Altschmied, L., and Graner, A. (2006). Genetic mapping and BAC assignment of EST-derived SSR markers shows non-uniform distribution of genes in the barley genome. Theor. Appl. Genet. 113:239-250.

Velasco, R., Zharkikh, A., Troggio, M., Cartwright, D.A., Cestaro, A., Pruss, D., Pindo, M., FitzGerald, L.M., Vezzulli, S., Reid, J., Malacarne, G., Iliev, D., Coppola, G., Wardell, B., Micheletti, D., Macalma, T., Facci, M., Mitchell, J.T., Perazzolli, M., Eldredge, G., Gatto, P., Oyzerski, R., Moretto, M., Gutin, N., Stefanini, M., Chen, Y., Segala, C., Davenport, C., Dematte, L., Mraz, A., Battilana, J., Stormo, K., Costa, F., Tao, Q., Si-Ammour, A., Harkins, T., Lackey, A., Perbost, C., Taillon, B., Stella, A., Solovyev, V., Fawcett, J.A., Sterck, L., Vandepoele, K., Grando, S.M., Toppo, S., Moser, C., Lanchbury, J., Bogden, R., Skolnick, M., Sgaramella, V., Bhatnagar, S.K., Fontana, P., Gutin, A., Van de Peer, Y., Salamini, F., and Viola, R. (2007). A High Quality Draft Consensus Sequence of the Genome of a Heterozygous Grapevine Variety. PLoS ONE 2:e1326.

Vera Ruiz, E.M., Soriano, J.M., Romero, C., Zhebentyayeva, T., Terol, J., Zuriaga, E., Llácer, G., Abbott, A.G., and Badenes, M.L. (2011). Narrowing down the apricot Plum pox virus resistance locus and comparative analysis with the peach genome syntenic region. Mol. Plant Pathol. 12:535-547.

Verlaan, M.G., Szinay, D., Hutton, S.F., De Jong, H., Kormelink, R., Visser, R.G.F., Scott, J.W., and Bai, Y. (2011). Chromosomal rearrangements between tomato and Solanum chilense hamper mapping and breeding of the TYLCV resistance gene Ty-1. Plant J. 68:1093-1103.

Vinson, J.P., Jaffe, D.B., O'Neill, K., Karlsson, E.K., Stange-Thomann, N., Anderson, S., Mesirov, J.P., Satoh, N., Satou, Y., Nusbaum, C., Birren, B., Galagan, J.E., and Lander. E.S. (2005). Assembly of polymorphic genomes: Algorithms and application to Ciona savignyi. Genome Res. 15:1127-1135.

Voorrips, R.E. (2002). MapChart: Software for the Graphical Presentation of Linkage Maps and QTLs. J. Hered. 93:77-78.

Vriesendorp, B., and Bakker, F.T. (2005). Reconstructing Patterns of Reticulate Evolution in *Angiosperms*: What Can We Do? Taxon 54:593-604.

Wada, H., Iwaya-Inoue, M., Akita, M., and Nonami, H. (2005). Hydraulic Conductance in Tepal Growth and Extension of Vase Life with Trehalose in Cut Tulip Flowers. J. Amer. Sot. Hort. Sci. 130:275-286.

Wall, P.K., Leebens-Mack, J., Chanderbali, A., Barakat, A., Wolcott, E., Liang, H., Landherr, L., Tomsho, L., Hu, Y., Carlson, J., Ma, H., Schuster, S., Soltis, D., Soltis, P., Altman, N., and dePamphilis, C. (2009). Comparison of next generation sequencing technologies for transcriptome characterization. BMC Genomics 10:347.

Wang, C.M., Liu, P., Yi, C., Gu, K., Sun, F., Li, L., Lo, L.C., Liu, X., Feng, F., Lin, G., Cao, S., Hong, Y., Yin, Z., and Yue, G.H. (2011). A First Generation Microsatellite- and SNP-Based Linkage Map of *Jatropha*. PLoS ONE 6:e23632.

Wei, H., Fu, Y., and Arora, R. (2005). Intron-flanking EST–PCR markers: from genetic marker development to gene structure analysis in *Rhododendron*. Theor. Appl. Genet. 111:1347-1356.

Wenzl, P., Carling, J., Kudrna, D., Jaccoud, D., Huttner, E., Kleinhofs, A., and Kilian, A. (2004). Diversity Arrays Technology (DArT) for whole-genome profiling of barley. Proc. Natl Acad. Sci. USA 101:9915-9920.

Wheat, C. (2010). Rapidly developing functional genomics in ecological model systems via 454 transcriptome sequencing. Genetica 138:433-451.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.-J., Makhijani, V., Roth, G.T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C.L., Irzyk, G.P., Lupski, J.R., Chinault, C., Song, X.-z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D.M., Margulies, M., Weinstock, G.M., Gibbs, R.A., and Rothberg, J.M. (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-876.

Whitelaw, C.A., Barbazuk, W.B., Pertea, G., Chan, A.P., Cheung, F., Lee, Y., Zheng, L., van Heeringen, S., Karamycheva, S., Bennetzen, J.L., SanMiguel, P., Lakey, N., Bedell, J., Yuan, Y., Budiman, M.A., Resnick, A., Van Aken, S., Utterback, T., Riedmuller, S., Williams, M., Feldblyum, T., Schubert, K., Beachy, R., Fraser, C.M., and Quackenbush, J. (2003). Enrichment of Gene-Coding Sequences in Maize by Genome Filtration. Science 302:2118-2120.

Wiens, J.J. (2005). Can Incomplete Taxa Rescue Phylogenetic Analyses from Long-Branch Attraction? Syst. Biol. 54:731-742.

Wittenberg, A., Lee, T., Cayla, C., Kilian, A., Visser, R., and Schouten, H. (2005). Validation of the high-throughput marker technology DArT using the model plant *Arabidopsis thaliana*. Mol. Genet. Genomics 274:30-39.

Woodcock, H., and Stearn, W. (1950). Lilies of the world. Country Life, London. pp 431.

Wu, F., Mueller, L.A., Crouzillat, D., Pétiard, V., and Tanksley, S.D. (2006). Combining Bioinformatics and Phylogenetics to Identify Large Sets of Single-Copy Orthologous Genes (COSII) for Comparative, Evolutionary and Systematic Studies: A Test Case in the Euasterid Plant Clade. Genetics 174:1407-1420.

Wu, H., Ramanna, M.S., Arens, P., and Van Tuyl, J.M. (2011). Genome constitution of *Narcissus* variety, 'Tete-a-Tete', analysed through GISH and NBS profiling. Euphytica 181:285-292.

Xia, L., Peng, K., Yang, S., Wenzl, P., Carmen de Vicente, M., Fregene, M., and Kilian, A. (2005). DArT for high-throughput genotyping of Cassava (*Manihot esculenta*) and its wild relatives. Theor. Appl. Genet. 110:1092 - 1098.

Xian-Liang, S., Xue-Zhen, S., and Tian-Zhen, Z. (2006). Segregation distortion and its effect on genetic mapping in plants. Chin. J. Agr. Biotechnol. 3:163-169.

Yamagishi, M., Kishimoto, S., and Nakayama, M. (2010). Carotenoid composition and changes in expression of carotenoid biosynthetic genes in tepals of Asiatic hybrid lily. Plant Breed. 129:100-107.

Yan, J., Yang, X., Shah, T., Sánchez-Villeda, H., Li, J., Warburton, M., Zhou, Y., Crouch, J., and Xu, Y. (2010). High-throughput SNP genotyping with the GoldenGate assay in maize. Mol. Breed. 25:441-451.

Yang, Z. (1997). PAML: A program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555-556.

Yang, Z., and Dos Reis, M. (2011). Statistical properties of the branch-site test of positive selection. Mol. Biol. Evol. 28:1217-1228.

Yang, Z., Nielsen, R., and Goldman, N. (2009). In defense of statistical methods for detecting positive selection. Proc. Natl Acad. Sci. USA. 106:E95.

Yang, Z., Nielsen, R., Goldman, N., and Pedersen, A.-M.K. (2000). Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. Genetics 155:431-449.

Yu, G., Olsen, K.M., and Schaal, B.A. (2011). Molecular Evolution of the Endosperm Starch Synthesis Pathway Genes in Rice (*Oryza sativa* L.) and Its Wild Ancestor, O. rufipogon L. Mol. Biol. Evol. 28:659-671.

Zerbino, D., and Birney, E. (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821-829.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011). A Practical Comparison of *De Novo* Genome Assembly Software Tools for Next-Generation Sequencing Technologies. PLoS ONE 6:e17915.

Zhong, Y., and Ciafré, C. (2011). Role of ABA in ethylene-independent *Iris* flower senescence. International Conference on Food Engineering and Biotechnology, IPCBEE, ACSIT Press, Singapoore volume 9.

Zhou, L., Jang, J.C., Jones, T.L., and Sheen, J. (1998). Glucose and ethylene signal transduction crosstalk revealed by an *Arabidopsis* glucose-insensitive mutant. Proc. Natl Acad. Sci. USA. 95:10294-10299.

Zhou, S. (2007). Intergenomic recombination and introgression breeding in Longiflorum x Asiatic lilies: Wageningen Universiteit.ISBN 9789085046370, pp 110.

Zhu, Q., Zheng, X., Luo, J., Gaut, B.S., and Ge, S. (2007). Multilocus Analysis of Nucleotide Variation of *Oryza sativa* and Its Wild Relatives: Severe Bottleneck during Domestication of Rice. Mol. Biol. Evol. 24:875-888.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P., Matukumalli, L.K., Grimm, D.R., Hyatt, S.M., Fickus, E.W., Young, N.D., and Cregan, P.B. (2003). Single-Nucleotide Polymorphisms in Soybean. Genetics 163:1123-1134.

Zonneveld, B.J.M. (2009). The systematic value of nuclear genome size for "all" species of *Tulipa* L. (*Liliaceae*). Plant Syst. Evol. 281:217-245.

Summary

Lily (*Lilium* L.) is a perennial bulbous ornamental, belonging to subclass *Monocotyledonae* and family *Liliaceae*. Lily, according to statistics of Dutch auctions, is the fifth most important cut flower and the second in flower bulbs based on acreage. This species has been extensively used for cytogenetic studies, but molecular genetic studies are limited. The heterogenic nature and the very complex and huge genome (36 Gb) of lily might be the reason for this. To improve the efficiency of breeding and selection in this species, and set up the basis for genetic studies in *Lilium*, genomic resources are needed.

Next generation sequencing (NGS) technology (454 pyro-sequencing) was used to sequence the transcriptomes (RNA-seq) of four lily cultivars: 'Connecticut King', 'White Fox', 'Star Gazer', and Trumpet that belong to the four most important hybrid groups: Asiatic, Longiflorum, Oriental, and Trumpet respectively. Successfully, 52,172 unigenes with an average length of 555 bp were developed and used for a wide range of genetic and genomic studies: SNP marker identification for genetic mapping, gene annotation, and comparative genomic studies.

Combining NGS with SNP genotyping techniques to accelerate genetic studies is of considerable interest in different species. In this study, thousands of SNPs out of the 52,172 lily unigenes were identified. Genotyping technique KASPar (KBiosciences competitive Allele Specific PCR) was used to genotype two lily mapping populations: 'LA (*L. longiflorum* 'White Fox' x Asiatic hybrid 'Connecticut King') and AA ('Connecticut King' x 'Orlito') using 225 SNP markers selected from 'Connecticut King' unigenes. Genotyping success rate was 75.5% (170 SNP markers worked), polymorphic SNP rate was 45% (102 SNP markers), and mapped SNP marker rate was 42% (94 SNP mapped) in LA population and 38% (85 SNP mapped) in AA population. Thus, we validated a subset of the putative SNP makers and showed the usability of this type of markers to improve genetic maps for complex genomes like that of lily.

The SNP markers together with the available AFLP (amplified fragment length polymorphisms), DArT (diversity arrays technology), and NBS (nucleotide binding site) markers were used to build reference genetic maps for these two lily populations. These maps represent the first reasonably saturated maps that cover 89% of the lily genome with an average marker density of one marker per 4 cM. The availability of more SNP markers for genotyping, opens the door for further enriching these genetic maps and thus improve the marker density.

The genetic maps were used to map and understand the genetic of several horticultural traits in *Lilium*. *Fusarium oxysporum* and lily mottle virus (LMoV) are considered as very serious diseases in *Lilium* and as such present important targets for breeding. Six putative QTLs (quantitive trait loci) were identified for *Fusarium* resistance in AA population, from which QTL1 was the strongest (explains ~25 % of phenotypic variation). In LA population, QTL1 was also confirmed. Thus, QTL1 is a strong and reliable QTL in both populations and it can be used

to develop markers for most of the *Fusarium* resistance for molecular assisted breeding (MAB) applications. The LMoV was mapped as a marker on the AA genetic maps, however, no close markers to this trait (*i.e.* distance of the closest marker to LMoV was 9 cM) were identified yet. Several ornamental traits: lily flower color 'carotene' (*LFCc*), flower spots (*lfs*), stem color (*LSC*), antherless phenotype (*lal*), and flower direction (up-side facing, *lfd*) were phenotyped and mapped. Some of these traits showed to be recessive traits (spots, antherless, and flower direction) and controlled by a single gene. Developing markers for recessive traits is valuable since such markers allow the identification of suitable breeding parents so the presence of the recessive trait can be either enhanced or repressed. A more complex trait is flower longevity because it is a function of: the number of buds per inflorescence, the expansion and opening of the buds, the life span of the individual flowers, and also the life-span of the leaves. Moreover, senescence in *Lilium* is ethylene-insensitive and the regulator(s) of its vase life is not known yet. Our study showed that vase life of individual lily flowers increased significantly by the exogenous application of sugars. Abscisic acid (ABA) level, furthermore, increased dramatically in lily flowers at senescence compared with anthesis. This indicates that ABA might be the main regulator of vase life in lily. However, more experiments should be conducted to prove this conclusion.

The genomic resources developed for lily together with genomic resources developed for *Tulipa* L., in the same way, offered a valuable source of information to conduct comparative genomic studies within and between these two genera. We initiated the first step towards linking molecular genetic maps of *Lilium* and *Tulipa* using transcriptome sequences generated by 454 pyro-sequencing. Orthologous genes between lily and tulip were identified (10,913 unigenes) based on sequence data of four lily cultivars and five tulip cultivars. Next, common SNP and EST-SSR markers between the parents of lily mapping populations (AA and LA population) and the parents of tulip mapping population ('Kees Nelis' (*T. gesneriana*) x 'Cantata' (*T. fosteriana*)) based on these orthologous sequences were generated. A total of 229 common SNP and 140 common EST-SSR markers were identified. Genotyping and mapping these markers in the populations of both genera will link the genetic maps of *Lilium* and *Tulipa* and thus allow insight into the preservation of gene order, structure, and 'putative' functional homology in addition to evolutionary processes.

Also, these genomic resources can be used to increase the resolution of, and support for, phylogenetic trees. We selected a set of orthologous genes of *Lilium* (19 genes, 11,766 bp containing 433 polymorphic sites), of *Tulipa* (20 genes, 10,347 bp containing 216 polymorphic sites), and of the orthologous genes between the two genera (7 genes, 5,790 bp containing 587 polymorphic sites). These sets are uniquely present in the sequences and informative in estimating the genetic divergence of the two genera, thus they can be used to genotypes more species per genera to build genera and maybe family trees later on. The nucleotide polymorphism rate of *Lilium* was twice as high as that of *Tulipa*, on average one substitution per 26 bp for

Lilium compared with one substitution per 48 bp for *Tulipa*. NGS provide a valuable source for large numbers of phylogenetic informative substitutions that might revolutionize the phylogenetic, population genetic, and biodiversity studies. However, the use of bi-allelic information from multiple loci in phylogenetic studies is still challenging and it needs to be studied further.

Moreover, having such high numbers of sequence data, allows us to test some evolutionary hypotheses such as positive selection: selection during domestication/breeding processes might be imprinted in the species genome, which can be examined based on omega (dn/ds) values. The higher the omega value the stronger the indication of positive selection. Positive selection was recorded in *Lilium* and *Tulipa* genomes when this small subset of gene contigs (46) of the two genera was tested. Our hypothesis could not be confirmed, however to draw final conclusions on this matter, omega values for many more genes of the two genera have to be measured.

Finally, a wealth of putative molecular markers (SNPs and SSRs) has become available that can have direct applications for breeding in these genera. SNP markers are important since they are user friendly, efficient, transferable, and co-dominant markers. Applying high throughput genotyping technology to genotype the two lily populations improved the coverage of the two genetic maps. Also, genotyping the same SNP markers in the two populations facilitated the comparisons between the linkage groups of the two populations and will allow the construction of a consensus map. Consequently, exchange of genetic knowledge (mainly QTLs) between the two populations will be easier. The thousands of SNPs identified in the genome of the four lily cultivars opens the door for combining the current linkage mapping studies with association studies which will have a direct impact on improving the resolution of mapping and on MAB applications in *Lilium*.

Samenvatting

Lelie *(Lilium* L.) is een meerjarig bolgewas, behorend tot de subklasse *Monocotyledonae* en familie *Liliaceae*. Lelie is volgens de Nederlandse statistieken de nummer vijf van de belangrijkste snijbloemen en nummer twee wat betreft het areaal bolgewassen. De soort is uitgebreid gebruikt voor cytogenetische studies, maar zeer beperkt in moleculair genetisch onderzoek. Dit zal het gevolg zijn van het heterogene karakter en het erg complexe en grote genoom (36 Gb) van lelie. Om efficiënter te kunnen veredelen en selecteren, en voor het opzetten van genetische studies in lelie, zijn genomische hulpmiddelen in de vorm van sequenties en merkers een vereiste.

Met behulp van de "next generation sequencing" (NGS) technologie (454 pyro-sequencing) werd het transcriptoom (RNA) gesequenced van vier lelie cultivars: 'Connecticut King', 'White Fox', 'Star Gazer', en een Trompet-selectie, die behoren tot de vier belangrijkste hybride groepen nl. de Aziaten, Longiflorums, Orientals en Trumpets. Met succes werden 52,175 unieke genen (unigenen) met een gemiddelde lengte van 555 basenparen ontwikkeld en gebruikt voor een reeks van genetische en genomische studies: DNA polymorfisme studies (SNP) door middel van merker identificatie voor de genetische kartering, gen annotatie en vergelijkend genomisch onderzoek.

Het is van belang voor de versnelling van genetische studies bij diverse species om de modernste sequencing technieken (aangeduid als NGS) met SNP-genotypering te combineren. In deze studie werden in de 52,175 unigenen duizenden SNPs geïdentificeerd. De genotyperings techniek KASPar (KBiosciences competitive Allele Specific PCR) werd gebruikt voor het genotyperen van twee kruisingspopulaties: LA (*L. longiflorum* 'White Fox' x Aziatische hybride 'Connecticut King') en AA ('Connecticut King' x 'Orlito') gebruikmakend van 225 SNP merkers geselecteerd in 'Connecticut King' unigenen. Met succes werd 75.5% (170 SNP-merkers werkten) van de SNPs gegenotypeerd, waarvan 45% polymorf was (102 SNP merkers). Hiervan kon 42% worden gekarteerd in de LA populatie (94 SNPs) en 38% in de AA populatie (85 SNPs). Op deze wijze werd een deel van de mogelijke SNP merkers gevalideerd en werd de bruikbaarheid van dit type merkers voor het verbeteren van genetische kaarten bij complexe genomen zoals lelie aangetoond.

De SNP merkers werden tezamen met de beschikbare AFLP (amplified fragment length polymorphisms), DArT (diversity arrays technology) en NBS (nucleotide binding site) merkers gebruikt om voor deze twee lelie populaties genetische referentie kaarten te construeren. Deze kaarten vertegenwoordigen de eerste redelijk verzadigde kaarten die 89% van het lelie genoom omvatten met een gemiddelde merker dichtheid van 4 cM per merker. De beschikbaarheid van meer SNP merkers maakt het mogelijk om deze genetische kaarten verder uit te breiden om daarmee de merker dichtheid te verbeteren.

De genetische kaarten van lelie werden gebruikt om verschillende tuinbouwkundige eigenschappen te karteren en om de genetica ervan te begrijpen. *Fusarium oxysporum* en het lily mottle virus (LMoV) behoren tot de belangrijkste ziekten in lelie en zijn als zodanig belangrijke veredelingsdoelen. Zes vermeende QTLs (quantitive trait loci) voor *Fusarium* resistentie werden geïdentificeerd in de AA-populatie waarvan QTL1 de sterkste is (verklaart 25% van de fenotypische variatie). Ook in de LA populatie werd QTL1 bevestigd. QTL1 is dus een sterke en betrouwbare QTL in beide populaties en kan gebruikt worden om merkers te ontwikkelen voor *Fusarium* resistentie bij de toepassing van moleculair gestuurde veredeling (ook wel aangeduid met MAB). De LMoV resistentie werd als merker gekarteerd op de genetische kaarten van de AA-populatie, maar er werden tot nu toe geen nauw gekoppelde merkers geïdentificeerd (de merker met kleinste afstand tot LMoV bevond zich op 9cM). Verschillende sierteelt kenmerken: lelie bloemkleur 'caroteen' *(LFCc*), bloemkleurspikkels (*lfs*), bloemsteelkleur (*LSC*), mannelijke steriliteit (*lal*), en bloem stand (rechtop, *lfd*) werden gefenotypeerd en gekarteerd. Enkele van deze recessieve eigenschappen (spikkels, mannelijke steriliteit en bloemstand) worden door een enkel gen bepaald. Ontwikkeling van merkers voor recessieve genen is waardevol omdat zulke merkers het mogelijk maken om de juiste kruisingsouders te identificeren zodat de aanwezigheid van de recessieve eigenschap al dan niet kan worden vermeden dan wel gestimuleerd. De houdbaarheid van de bloem is een meer complexe eigenschap omdat deze bepaald wordt door het aantal bloemen per tak, de groei en ontwikkeling, de opening van de knoppen en de levensduur van de individuele bloemen en de bladeren. Bovendien is veroudering in lelie ethyleen-ongevoelig en kennis over de regulering van het vaasleven van lelie is nog onvoldoende. Onze studie toonde aan dat het vaasleven van individuele leliebloemen aanzienlijk verhoogd werd onder invloed van extern toegevoegde suikers. Het abscissine zuur (ABA) gehalte bleek bovendien aanzienlijk toe te nemen bij verouderende bloemen ten opzichte van vers geopende bloemen. Dit wijst erop dat vooral ABA de regulering van het vaasleven van lelie bepaald. Er moet echter meer onderzoek verricht worden om deze conclusie te bevestigen.

De genomische hulpmiddelen voor lelie, die tezamen met die voor tulp (*Tulipa*) op dezelfde wijze zijn ontwikkeld, bieden een belangrijke bron van informatie voor de uitvoering van vergelijkend genomisch onderzoek binnen en tussen deze twee geslachten. We hebben een eerste stap gezet naar het koppelen van de moleculair genetische kaarten van lelie en tulp door gebruik te maken van transcriptoom sequenties verkregen door '454 pyro-sequencing'. Orthologe genen tussen lelie en tulp werden geïdentificeerd (10,913 unigenen) gebaseerd op sequentie gegevens van vier lelie en vijf tulpen cultivars. Vervolgens werden gemeenschappelijke SNP en EST-SSR merkers tussen de ouders van de lelie merker populaties (AA en LA) en de ouders van de tulpen merker populatie ('Kees Nelis' (*T. gesneriana*) x 'Cantata' (*T. fosteriana*)) gebaseerd op deze orthologe sequenties gegenereerd. In totaal werden 229 gemeenschappelijke SNP en 140 gemeenschappelijke EST-SSR merkers geïdentificeerd. Door genotypering en kartering van deze merkers in de populaties van beide geslachten kunnen de genetische kaarten van lelie en tulp gekoppeld worden en inzicht worden verschaft in de conservering van de genenvolgorde, de structuur en 'mogelijke' functionele homologie evenals in evolutionaire processen.

Ook kunnen deze genomische data worden gebruikt om de resolutie van en ondersteuning voor fylogenetische stambomen te verhogen. We selecteerden een set van orthologe genen in lelie (19 genen, 11,766 bp waarin 433 polymorfe plaatsen), en in tulp (20 genen, 10,347 bp met 216 polymorfe plaatsen), en van de orthologe genen tussen de twee geslachten (7 genen, 5,790 bp met 587 polymorfe plaatsen). Deze sets zijn uniek aanwezig in de sequenties en zijn informatief ten aanzien van het schatten van de genetische divergentie van de twee geslachten, en kunnen worden gebruikt om meer species per geslacht te genotyperen en om later een geslachts- of familie stamboom te construeren. Het niveau van nucleotide polymorfismen bij lelie was twee keer zo hoog als bij tulp, gemiddeld één substitutie per 26 bp voor lelie tegen 48 bp voor tulp. NGS verschaft een waardevolle bron van grote aantallen fylogenetisch informatieve substituties die een revolutie kunnen betekenen voor fylogenetisch, populatie genetisch en biodiversiteitsonderzoek. Het gebruik van bi-allelische informatie van multipele loci in fylogenetische studies is echter een uitdaging en vereist nader onderzoek.

Zulke grote hoeveelheden sequentie data maakt het bovendien mogelijk om enkele evolutionaire hypothese te toetsen zoals positieve selectie: selectie tijdens het domesticatie/veredelingsproces zou in het genoom van de soort vastgelegd kunnen zijn, wat onderzocht kan worden op basis van omega (dn/ds) waarden. Hoe hoger de omega waarde des te sterker de aanwijzing voor positieve selectie. Positieve selectie werd in lelie en tulp waar genomen in de kleine set van gen contigs (46) van de twee geslachten. Onze hypothese kon niet bevestigd worden, omdat voor een uiteindelijke conclusie de omega waarden van veel meer genen van de twee geslachten gemeten moeten worden.

Ten slotte, er is een schat aan mogelijke moleculaire merkers (SNPs en SSRs) beschikbaar gekomen welke direct toegepast kunnen worden bij de veredeling van deze geslachten. SNP merkers zijn belangrijk omdat het gebruikersvriendelijke, efficiënte, overdraagbare en co-dominante merkers zijn. Door toepassing van een high-throughput genotyperings technologie werden twee lelie populaties gegenotypeerd en werd de dekking van de twee genetische kaarten verbeterd. Door de genotypering van deze SNP merkers in de twee populaties werd een vergelijking mogelijk tussen de koppelingsgroepen van de twee populaties en wordt de constructie van een consensus kaart mogelijk. Dit heeft tot gevolg dat uitwisseling van genetische kennis (vooral QTLs) tussen de populaties gemakkelijker zal worden. De duizenden SNPs die in het genoom van de vier lelie cultivars geïdentificeerd zijn, maakt het mogelijk om de huidige koppelingsstudies te combineren met associatiestudies wat een directe impact zal hebben op de verbetering van de resolutie van de kartering en op de MAB toepassingen in lelie.

Acknowledgments

It is an achievement of almost four years of hard work and support of a bunch of very nice people. With pleasure and satisfaction I am going to say few words not only to acknowledge but also to express my gratitude to all those who helped me to face all the challenges and shared the happiness with me during my stay in the Netherlands.

This story started when I received an invitation letter form **Dr. Jaap van Tuyl** accepting me to accomplish my PhD thesis at his group. Together with a scholarship from **Damascus University**, I started my journey with research at Wageningen University and Research Centre. Jaap, I will be always grateful to you and to your wife "**Nolly**" for opening your house to me and treating me as a daughter, I will not forget your words "I am like your father, you should listen to me!" Thanks a lot Jaap; you are a very good group leader and very supportive supervisor.

I would like to extend my thanks to my promoter **Prof. Richard Visser.** Although you are so busy, you still find time to discuss my progress, read my manuscripts, and answer my emails quite fast. Our meetings were always useful and to the point. Thanks Richard for being so clear, direct, and supportive.

Many thanks go to my daily supervisor and co-promoter **Dr. Paul Arens**. Working with new topics and in new fields of research was always a serious challenge, but with your ability to ask always the right questions we could make it. Thanks for all your in depth scientific comments and the critical reading of my manuscripts that helped me always to improve my work. Thanks for opening your door always to receive my 5 min question!! Thanks Paul a lot for your understanding and support.

I would like to acknowledge my group "**Ornamental group**" for their nice company during Monday morning meetings and Excursions under Jaap lead. Here, I would like to express my gratitude to "**Alex van Silfhout**" for his presence in this group. He could always smile, talk, and build friendship with each group member. Thanks Alex for making a lovely environment. Thanks for being helpful for every person in our group, ready to listen and support, and for your kindness. I would like to expend my thanks to **Dr. Agnes Ciolakowki-Marasek** for her friendship and for the wonderful time I spend with her family. Also, I would like to thank the students **Jelle, Han, Ellena,** and **Hwang** for their input.

I would like to thanks **Damascus Univeristy** for offering me a scholarship to do my PhD, and to the **Dutch lily breeding companies** (De Jong Lelies BV, Royal Van Zanten BV, Vletter and Den Haan BV, Mak Breeding BV, Marklily BV, World Breeding BV, Vanden BosBreeding BV, and Steenvoorden BV together) together with **TTI Grenomics** organization for funding this project.

The secretary committee were part of our daily work environment, with their smile things were done easily and smoothly. Thanks **Annie**, **Lettey**, **Mariame Gada**, **Janneke**, and **Nicole** for your help in administrative and financial issues.

I would like to express my sincere thanks to **Dr. Eric van de Weg**. Eric, I am honored to know you, and indebted for your concern and your time which you spend listening to me. You could always support, advice and help me, thanks a lot. I would like also to thank **Dr. Sjaak van Heusden** for being very positive and supportive. Many thanks to **Dr. Rene Smulders** and **Dr. Freek Bakker** for their scientific contribution and guidance in plant systematic field, without your support I could not complete this chapter. Thanks a lot to **Dr. Theo Born** for helping me submitting the data to gene bank. **Dr. Chris Maliepaard**, thanks for your statistical feedback. **Dr. Henk Schouten** and **Dr. Herman van Eck** thanks for the nice discussion we had during coffee time.

Special thanks to my external supervisor **Prof. Hans de Jong**, Hans it was always so pleasant and beneficial to listen to your advices and your scientific experience. Thanks for inviting me to the Chromosome conference, which was so kind of you. I would like also to thank Prof. **Fawaz Alazmeh** who opened the door for me to fly and build my career. Thanks Prof. Fawaz for your constant support and your concern. With your advice, I was very confident about my choices.

My acknowledgments to **Martijn van Kawwen**, **Danny Esselink,** and **Thomas van Gurp** for being there to help me solve software problems, analyze data, and understand bioinformatics topics. Thanks to all of you. Also, I would like to thank **Elma Salentijn** and **Jan Schaart** for their delicate kindness in receiving me at their office. We could hold on (together with Alex) long discussions about all kind of topics. That was very enjoyable and relaxing time, thanks.

During PhD period, I got to know very special and honest friends that I will always remember with a lot of respect. **Xingfeng Huang** and **Nadeem Khan** (my officemates), and **Lidia Rebelo Lima** (my housemate) you have been wonderful and great company. I learnt from you how to cope with working environment and keep smile even when I am under fire!! Lidia, thanks for your friendship, for honesty, and for sharing life experience with me. I will never forget all our talks discussing and sharing all types of stories that passed through in our life. Thanks to all of you my friends, I will never forget your smiles☺.

Thanks to all my international friends and colleagues who made my stay in Wageningen unforgettable. Thanks for the very nice company, nice gathering, and useful discussion. It is my pleasure to know you all and I really wish to keep in touch. Thanks my Iranian friends: **Nasim**, **Leila**, and **Shiva y**ou have been great friends, I enjoyed a lot our outing and the fun we had together. Nasim, maybe it is very strange to thank for sharing stress. But honestly, I would like to thank you for sharing all the stress of the last year of our PhDs, mainly at Christmas and New Year 2012 when all our friends were on holiday. Thus, we celebrated 2012 New Year's Eve just both of us, and it was very nice time thanks! **Paola Pollegioni** very special thanks; Rome was great with your company! **Hulya** thanks for sharing nice moments with me, **Luis** the great dancer, **Mirjana** the sweet sportive girl, **Anna** the direct and clear lady, **Maria** the lovely friend, **Thijs** and **Yusuf** for hard discussion and fun, **Animesh** 'sharing is caring', and many more

friends: **Songlin**, **Feras**, **Nadia**, **Lisa**, **Freddy**, **Marian**, **Wei**, **Ram**, **Sabaz**, **Pierre**, **Anitha**, **Paula**, **Cesar**, **Giulia**, **Virginia**, **Brigitte**, **Anoma**, **Björn D'hoop, Bjorn Kloosterman, Nicols**, **Alireza**, **Peter**, **Ningwen**, **Mathieu**, **Paweena**, **Efstathios, Rafael, Christos, Marteen, Olympia, Elena, Benyamin**, **Sohail**, **Thierry**, **Stefano**, **Zheng**….and many more of nice fiends…

I would like to devote my acknowledgments to my Syrian friends: **Eman Rustum**, **Nuha Salum**, and **Fatima Hatem** you are my prodigious friends, thanks for being there to listen and share with me my silly stories, thanks for trusting and believing in me. Fatima, we grow up together and I hope to keep our relationship as long as we are alive. Also, my thanks go to **Manal**, **Hayan**, **Aroub**, **Linda**, **Suha**, and **Saja** for your nice friendship. Special thanks to **Dr. Safaa**, and **Dr. Ramzi** for your hospitality in France, it was a great time that I will never forget. Safaa, thanks for your constant advice and concern. **Dr. Yamen** and **Dareen**, the very lovey couple, thanks for opening your house for me, I had great time with you.

I would like to thank **Alex van Silfhout** and **Mirjana Vukosavljev** for being my paranymprhs and share with me my special day. Looking forward to have a great day!

Deepest gratitude and thanks to **my family**. Traveling abroad was not common in my family. It was very hard for you to accept the idea that I will be far for four years. You have been counting the days for me to finish with endless worry of being alone in a foreign country. I would like to thank my parents for believing in me and for your continuous encouragement, without that I would never have been able to reach this point. Thanks to my brothers: **Ayham**, **Alwan**, and **Mothanna**, to my sisters: **Khawla** and **Azza**, to my brothers and sister in low: **Ayad**, **Ranim**, and **Wael**, and thanks to the next generation of my family nieces and nephew: **Juddy**, **Zynab**, **Gana**, **Zakarya**, **Al-Zahraa**, and **Al-Battoul**. Your presence makes my life soulful and colourful.

At the end, I would like to extend my thanks to you '**All**' who helped me in one way or another and I could not mention their names.

Arwa Shahin

Wageningen UR

About the author

Arwa Shahin was born on June 1st, 1980 in Tartus, Syria. She received her BSc in Agricultural Engineering in 2003 and Diploma in Horticulture Science in 2004 from Damascus University, Syria. Under TEMPUS programme (collaboration project between Syrian goverment and European Union) she accomplished her MSc degree in Biotechnology in 2007 that was done at Damascus University and Ghent Univerity, Belgium. She worked as a research assistant from May 2004 till June 2005 at the National Commission for Biotechnology, Syria, and from July 2005 till December 2008 as academic instructor at Damascus University, Syria. In January 2008 she started her PhD program at the laboratory of Plant Breeding, Wageningen University and Research Centre (WUR), the Netherlands. This thesis presents the outcome of her four years PhD research work on "Development of Genomic Resources for Ornamental lilies (*Lilium* L.).

List of Publications:

- Shahin A, Arens P, van Heusden S, Van Tuyl JM: **Conversion of molecular markers linked to *Fusarium* and Virus resistance in Asiatic lily hybrids**. Acta Hort (ISHS) 2009, **836**:131-136.
- Shahin A, Arens P, Van Heusden AW, Van Der Linden G, Van Kaauwen M, Khan N, Schouten HJ, Van De Weg WE, Visser RGF, Van Tuyl JM: **Genetic mapping in *Lilium*: mapping of major genes and quantitative trait loci for several ornamental traits and disease resistances**. Plant Breeding 2011, **130**(3):372-382.
- Shahin A, van Gurp T, Peters SA, Visser RGF, van Tuyl JM, Arens P: **SNP markers retrieval for a non-model species: a practical approach**. BMC Research Notes 2012, 5:79.
- Tuyl JM, Arens P, Ramanna MS, Shahin A, Khan N, Xie S, Marasek-Ciolakowska A, Lim K-B, Barba-Gonzalez R, Kole C: ***Lilium*.** In: Chittaranjan Kole (ed). **Wild Crop Relatives: Genomic and Breeding Resources**. In.: Springer Berlin Heidelberg; 2011: 161-183.
- Shahin A, Al-Marrei K: **Pre-selection of lime tolerant varieties of *Pyrus communis***. Al-Bassel Journal 2006, Syria.
- Shahin A, Van Kaauwen M, Esselink D, Van Tuyl JM, Visser RGF, Arens P: **Generation and analysis of expressed sequence tags in the extreme large genomes *Lilium* L. and *Tulipa* L**. Submitted.
- Shahin A, Van Tuyl JM, Visser RGF, Arens P: **Genotyping and mapping of SNP markers in *Lilium* L**. To be submitted.
- Shahin A, Smulders MJM, Bakker FT, Van Tuyl JM, Visser RGF, Arens P: **Using *Lilium* L. and *Tulipa* L. high-throughput sequencing data for estimating genetic distances and positive selection**. To be submitted.
- Shahin A, Van Silfhout A, Verstappen F, Bouwmeester H, Van Tuyl JM, Visser RGF, Arens P. **Towards a better understanding of the vase life of lily flowers**. Submitted.
- Smulders MJM, Vukosavljev M, Shahin A**,** Van De Weg WE, Arens P: **High throughput marker development and application in horticultural crops**. Acta Hort (ISHS). In press.

Education Statement of the Graduate School

Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: **Arwa Shahin**

Date: **19 June 2012**

Group: **Laboratory of Plant Breeding, Wageningen University & Research Centre**

1) Start-up phase	*date*
► **First presentation of your project**	
Molecular cytogenetic approaches for transferring genes for resistance in Lily and tulip	Dec 01, 2008
► **Writing or rewriting a project proposal**	
Molecular cytogenetic approaches for transferring genes for resistance in Lily and tulip	Jun-Jul 2008
► **Writing a review or book chapter**	
Lilium: In Wild Crop Relatives: Genomic and Breeding Resources, Plantation and Ornamental Crops, Springer-Verlag (2011)	Sep 2009
High throughput marker development and application in horticultural crops, submitted	2012
► **MSc courses**	
Breeding for resistance and quality (30306)	Jul 02, 2008
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	***17,5 credits****

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD student days, Leiden University, Netherlands	Mar 26, 2009
EPS PhD student days, Utrecht University, Netherlands	Jun 01, 2010
2nd Joint Retreat of PhD Students in Experimental Plant Science, Max Panck, Germany	Apr 15-17, 2010
EPS PhD student days, Wageningen University, Netherlands	May 20, 2011
► **EPS theme symposia**	
EPS theme 4 ' Genome Plasticity' Wageningen University	Dec 12, 2008
EPS theme 4 ' Genome Plasticity' Wageningen University	Dec 10, 2010
EPS theme 4 ' Genome Plasticity' Wageningen University	Dec 09, 2011
► **NWO Lunteren days and other National Platforms**	
NWO-ALW Experimental Plant Science, Lunteren, Netherlands	Apr 06-07, 2009
NWO-ALW Experimental Plant Science, Lunteren, Netherlands	Apr 19-20, 2010
NWO-ALW Experimental Plant Science, Lunteren, Netherlands	Apr 04-05, 2011
► **Seminars (series), workshops and symposia (highly recommended)**	
European Flying Seminar Prof. dr. Simon Gilroy, Title: 'How do plants feel? Mechanical Signaling in *Arabidopsis*'	May 19, 2008
Raising the BAR for *Arabidopsis* Research: Using Large-scale Data Sets for Hypothesis Generation'	Jun 13, 2008
Mechanism and function of active DNA demethylation in *Arabidopsis*	Nov 03, 2008
Partitioning the genome: mechanisms that ensure accurate chromosome segregation in cell division' by Dr. Michael Lampson	Feb 01, 2008
Tomato innate immunity to root-knot nematodes and aphids' by Dr. Isgouhi Kaloshian	May 14, 2009
The molecular regulation of seed dormancy' by Wim Soppe	Oct 20, 2009

	'The molecular dialogue between pathogens and plants' by Pierre de Wit	Nov 10,2009
	'Statistical modeling of genotype to phenotype relations' by Fred van Eeuwijk	Nov 10,2009
	SNP detection and next gen sequencing data' by Jack Leunissen	Apr 06, 2010
	WIAS Symposium 'Genomics and Animal Breeding', Wageningen, Netherlands	Jan 21, 2011
	Plant Research Day (Plant Breeding)	Jun 17, 2008
	Plant Research Day (Plant Breeding)	Mar 08, 2011
	Plant Research Day (Plant Breeding)	2010
	Plant Research Day (Plant Breeding)	Feb 28, 2012
	Workshop EU-SPICY "Bioinformatics, statistical genetics and genomics", Wageningen, Netherlands	Mar 08, 2012
►	**Seminar plus**	
	PhD discussion 'Plant Soil Interactions: the design of experiments and Genotype x Environment interactions' by Prof Fred van Eeuwijk	Mar 15, 2011
►	**International Symposia and Congresses**	
	23th EUCARPIA Symposium, Leiden 2009 (Colorful Breeding and Genetics), Netherlands	Aug 31-Sep 04, 2009
	28th International Horticultural congress, Lisbon, Portugal	Aug 22-27, 2010
	2nd International Symposium on genus *Lilium*, Pescia (PT), Italy	Aug 30-Sep 03, 2010
	TTI green genetics Symposium, Netherlands	Sep 22, 2010
	CBSG Technology Symposium "Advance in life-science Technologies", Netherlands	Nov 25, 2010
	18th International Chromosome Conference, Manchester, UK	Aug 29-Sep 02, 2011
	TTI green genetics Symposium, Netherlands	Sep 21, 2011
	The XIth International Symposium on Flower Bulbs and Herbaceous Perennials, Turkey	Mar 28-Apr 01, 2012
►	**Presentations**	
	Plant Breeding meeting (research day): Molecular cytogenetic approach for transferring genes of *Fusarium* resistance in lily (poster)	Jun 17,2008
	Presentation for Lily companies (Molecular assisted breeding for *Fusarium* and LMoV in lily)(oral)	Jun 24,2008
	Theme 4 (Genome Plasticity): Integrated Cytological and Molecular Map of Lily and Synteny to Tulip (oral)	Dec12, 2008
	Oral Presentation for Lily companies (Conversion markers linked to resistance in lily)(oral)	Dec18, 2008
	EUCARPIA Symposium 2009 (Conversion of Molecular Markers linked to *Fusarium* and Virus Resistance in Asiatic Lily Hybrids)(oral)	Sep 02, 2009
	Presentation for Japanese group (Niigata University) (Towards lily genome-wide understanding)(oral)	Mar 09, 2010
	2nd joint PhD retreat (Genetic Mapping in *Lilium*)(poster)	Apr 15-17, 2010
	Presentation for Lily Breeding Companies (Bridging the genomes of lily)(oral)	Jul 13, 2010
	28th International Horticultural congress) 2010- Lisbon (Development of EST Derived SNP Markers in *Lilium* and Tulip)(oral)	Aug 24, 2010
	28th International Horticultural congress) 2010- Lisbon (Using SNPs Markers for Resistance Breeding in *Lilium*)(poster)	Aug 26, 2010
	Lilium Symposium 2010- Italy (Mapping of *Fusarium oxysporum* resistance in *Lilium*)(oral)	Sep 03, 2010
	TTI Green Genetics 2010- Netherlands (Using SNPs Markers for Resistance Breeding in *Lilium*)(oral)	Sep 22, 2010
	TTI Green Genetics 2011- Netherlands (Development of SNPs Markers for Resistance Breeding in *Lilium*)(poster)	Sep 21, 2011
	Lily Breeding Companies & TTI management (Bridging the genomes of lily)(oral)	Dec 13, 2011
	XIth Intern. Symp. on Flower Bulbs and Herbaceous Perennials-Turkey (Genotyping and mapping of SNP markers in *Lilium*)(oral)	Mar 28, 2012
	XIth Intern. Symp. on Flower Bulbs and Herbaceous Perennials-Turkey (What controls flower longevity in *Lilium*)(poster)	Mar 28, 2012
►	**IAB interview**	Feb 17, 2011

►	**Excursions**	
	Visit Syngenta company	Sep 25, 2008
	Visit Lily and Tulip companies	May 14, 2009
	Visit Flower breeding companies/ institutes and auction	May 19, 2009
	Visit Flower breeding companies (Keukenhof, Florist, MarkLily, PPO, Remarkable Tulip)	May 11, 2011
	Visit Darthuizer company, Syngenta flower company, Mak company (Lily breeding company), and Keukenhof	May 18, 2011
	Visit Royal van Zanten flower company	Mar 07, 2012
Subtotal Scientific Exposure		***34,4 credits****

3) In-Depth Studies		*date*
►	**EPS courses or other PhD courses**	
	PhD Course 'Analysis of ~Omics data', Wageningen, Netherlands	Dec 08-11,2008
	PhD Workshop 'Plant Metabolomics', Wageningen, Netherlands	Apr 26-28, 2011
	PhD Course 'Current Trends in Phylogenetics', Wageningen, Netherlands	Oct 24-28, 2011
►	**Journal club**	
	Member of literature discussion group of Plant Breeding	2008-2011
►	**Individual research training**	
Subtotal In-Depth Studies		***6,6 credits****

4) Personal development		*date*
►	**Skill training courses**	
	PhD Competence Assessment	May 12, 2009
	Academic writing I	2009
	Academic writing II	Mar 03, 2010
	Techniques for Writing and Presenting a Scientific Paper	Apr 24, 2012
►	**Organization of PhD students day, course or conference**	
	Organizing the Biweekly Colloquium of Plant Breeding Department-group of breeding for growth, development and quality	Jan-Oct 2010
►	**Membership of Board, Committee or PhD council**	
Subtotal Personal Development		***4,5 credits****

TOTAL NUMBER OF CREDIT POINTS*	**63.0**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

Cover idea: By the author

Cover design: By Mani Rezaeian

mani.rezaeian@yahoo.com

www.Gishniz.com

Printing: Ipskamp Drukkers BV, Nijmegen, the Netherlands