Genetic Diversity of Proprietary Inbred Lines of Sunflower, Determined by Mapped SSR Markers and Total protein analysis

by

Tertia Elizabeth Erasmus

Submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy in

the School of Agricultural Science and Agribusiness

Faculty of Science, University of KwaZulu Natal

Pietermaritzburg

2008

Thesis Abstract

This study compared DNA based SSR markers with total seed protein markers, used to evaluate genetic diversity of sunflower. The multiplex-ability, cost effectiveness and applicability of microsatellites as molecular markers for a genetic diversity study were investigated and evaluated based on pedigree data of the sunflower germplasm. A solution for oil and fat interference in ultrathin iso-electric focusing gels was investigated, in order to make imaging and interpretation easier and clearer. Total protein analysis was utilized for the determination of genetic diversity on the same inbred material used for the DNA analysis. Finally a correlation is made between the data obtained on DNA vs Protein compared with phenotype and expected pedigree data.

A set of 73 SSR markers with known mapped positions were utilized to determine genetic similarity in a group of sunflower inbred lines. Cluster analysis of genetic similarity revealed an excellent correlation with the breeding background and source information obtained from breeders on all inbred lines used in this study. Cluster analysis gave a clear differentiation between B and R-lines, showing clearly defined heterotic groups of the proprietary set of inbred lines.

The most outstanding single-locus SSR markers in the set used for this study were identified and used as a core set. Multiplex assays were designed and optimized for the most cost and time effective method for rapid variety identification. The selected markers produced robust PCR products, amplified a single locus each, were polymorphic among the elite inbred lines and supplied a good, genome-wide framework of completely co-dominant, single-locus DNA markers for molecular breeding. The use of a fluorescent-tailed primer technique resulted in a considerable cost saving. Furthermore, the SSR markers can be multiplexed through optimization, in order to avoid undesirable primer-primer interactions and non-specific amplification.

First stage iso-electric focusing of total protein extracts were used to analyze sunflower looking at genetic purity and genetic variety verification on diverse sunflower germplasm. Severe visual interference was visible on most seed storage protein extracts of sunflower. This interference was visible as a distortion in the gel matrix on the anodal end of the gel, and caused important proteins to denature in the presence of heightened field strength and the absence of a uniform matrix. Adjustment of the extraction solutions removed this interference.

Total protein profiles were generated with the use ultrathin layer iso-electric focusing (UTLIEF) to assess the level of genetic diversity on the same set of sunflower lines used for the SSR analysis. Finally, the genetic diversity of the sunflower germplasm was analysed by comparing proteomic, genomic and pedigree data from the same germplasm. A total of 295 alleles were amplified with a set of 73 SSR markers with known mapped positions. These were utilized to determine the genetic relatedness of a group of B-lines and R-lines of sunflower. In parallel, a total of 68 protein bands were visualized using protein samples of two types of seed storage proteins derived from exactly the same sunflower lines. Cluster analysis clearly differentiated between the B-lines and R-lines, identifying defined heterotic groups of this proprietary set of lines. The comparison of DNA and protein data for the application of genetic diversity studies is analysed, as well as the general comparison on the use of the two different molecules as markers.

Declaration

I Tertia Elizabeth Erasmus declare that

(i) The research reported in this dissertation, except where otherwise indicated, is my original work.

(ii) This dissertation has not been submitted for any degree or examination at any other university.

(iii) This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons.

(iv) This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:

a) their words have been re-written but the general information attributed to them has been referenced;

b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced.

(v) Where I have reproduced a publication of which I am an author, co-author or editor, I have indicated in detail which part of the publication was actually written by myself alone and have fully referenced such publications.

(vi) This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the References sections.

Signed:…………………………….

Tertia Erasmus

Signed:…………………………….
Prof. M.D. Laing (Supervisor)

Table of Contents

Acknowledgements

My sincere thanks are extended to all the people who made this project possible:

My thanks go to my supervisor Professor M.D. Laing, for his consistent dedication and support throughout this study. His insight and refreshing approach to the project was encouraging. Thank you for all your support throughout all the challenges that I had to face in completing this study.

I am grateful to PANNAR (Pty) Ltd, for their contribution.

My thanks to Incotec Proteios South Africa for their understanding and time that they allowed me to complete this project. A special thank you to my brother Hannes Hattingh and his wife Tia Hattingh for their support.

A heartfelt thank you to my friends, Linda Chiazzari and Anne Skelton for the continuous encouragement and support,

Finally a sincere thanks to my dear family, my husband **Lourens Erasmus** and my children, **Jeane-Marie** and **Theo** for their patience and support even during times of my “absent minded” presence. In particular a sincere and intense thank you to my mother **Jeanetta Hattingh** and my departed father **Theo Hattingh** for all the support and encouragement through the years of study, they always shared this goal.

Foreword

This thesis is the product of some years of study and experience in protein and DNA analysis in a commercial environment. Chapter 1 reviews literature concerning Sunflower and the various markers used in this study. Chapter 2 report the results of the genetic diversity study based on the use of SSRs. Chapter 3 reports on a optimal core set of SSRs for a unique multiplex PCR strategy. Chapter 4 presents a solution for the visual interference common in sunflower protein gel electrophoresis that is often encountered in general PAGE. Chapter 5 covers the results of genetic analysis through the use of seed storage proteins and ultra-thin layer iso-electric focusing. Finally in Chapter 6 a correlation is discussed between genetic diversity data obtained from DNA, Protein and phenotypic data (though limited due to confidentiality issues) based on the pedigree data obtained on the inbred lines used in this study. Chapter 7 is an overview of the goals achieved and future research possibilities forthcoming from this project.

The chapters are written as discrete papers, in the “Dutch” style of thesis. This results in a level of duplication of references between chapters, and between the chapter abstract and the overall thesis abstract.

CHAPTER 1: A literature review on evaluating genetic diversity in sunflower (*Helianthus annuus* L.)

1.1 Introduction to sunflower

Sunflower (*Helianthus annuus* Linnaeus) is the second most important oilseed crop worldwide, after soybean (Paniego *et al*. 2002). Sunflower ranks among the first four oilseed crops in land area under production and seed production (Tang *et al*. 2002).

The genus *Helianthus* contains 12 annual and 37 perennial species (Hvarleva *et al*. 2007). The weedy, self-incompatible common sunflower is native to North America and was used by the native Americans before the colonization of the New World. According to Putnam *et al*. (1990), the sunflower was first introduced to Europe as an ornamental through Spain where by 1580, it was a common garden flower. It spread through the trade routes to Italy, Egypt, Afghanistan, India, China and Russia. In Russia selections for high oil content began in 1860, a process that has eventually increased the average oil content of sunflower seeds from 28% to 50%.

Sunflower has its value as an important crop because commercially available sunflower varieties contain 39 - 49% oil in their seed. The oil is considered a premium cooking oil because of its light colour, high level of unsaturated fatty acid, a general lack of linolenic acid, its bland flavour and a high smoke point. The primary fatty acids are oleic and linoleic (unsaturated) fatty acids, with small amounts of palmitic and stearic (saturated) fatty acids. High oleic sunflower oil (over 80% oleic acid) has a higher oxidated stability than most other cooking oils.

After oil extraction the residue of sunflower meal compares well with soymeal because it contains equal amounts of protein. Sunflower meal is higher in fibre, lower in energy value, lower in lysine but higher in methionine than soybean meal. It is commonly fed to ruminant animals, swine and poultry.

Sunflower oil also has industrial applications and has been used in certain paints, varnishes and plastics because of good semi-drying properties without colour modification that are associated with oils high in linolenic acid. The oil has been used as a pesticide carrier, and in the production of agrichemicals, surfactants, adhesives, plastics, fabric softeners, lubricants and coatings.

A key step in the conversion of sunflower into a major agricultural crop was the discovery of genes for cytoplasmic male sterility and male sterility restoration (Leclercq 1969), which allowed for the large scale production of hybrid seed. Male sterility is defined as the failure of plants to produce functional anthers, pollen, or male gametes, whereas female reproduction remains normal (Chen *et al*. 2006). Based in its inheritance or origin, male sterility may be divided into nuclear male sterility (NMS), also called genetic male sterility (GMS), and cytoplasmic male sterility (CMS). Both types of male sterility have been found in sunflower. NMS in sunflower was first reported in the Soviet Union by Kuptsov in 1934 (Chen *et al*. 2006). Cultivated sunflower are produced as hybrids, obtained by crossing a male sterile, female inbred line (A-line) with male fertile, restorer line (R-line). Pure seed of the male sterile A-line is produced by crossing it with an isogenic male fertile, maintainer line (B-line), since it cannot be self-pollinated. All hybrids use a single male sterile cytoplasm, derived from a wild annual sunflower, *Helianthus petiolaris* Nutt. This narrowing of the germplasm makes sunflower vulnerable to many insect and disease pests (Chen *et al*. 2006).

The most serious diseases of sunflower are caused by fungi, including rust, downy mildew, verticillium wilt, sclerotinia stalk and head rot, phoma black stem and leaf spot. Sclerotinia has the largest effect on crop yield. Resistance genes against rust, downy mildew, and verticillium wilt have been incorporated into improved germplasm. In a report on the status of sunflower in the US (Tang *et al*. 2003), it was noted that rust (*Puccinia helianthi*

Schewein), and downy mildew (*Plasmopara halstedii* (Farlow) Berlese & de Toni) have evolved with the crop, so that new races of these prolific and polycyclic pathogens are continuously evolving to match the progress of plant breeders, with the result that these pathogens can have a devastating effect on the crop, as can white rust caused by *Albugo tragopogonis* (DC.) Gray.

Cultivated and common sunflower are completely inter-fertile and are considered to be members of the same species. However, they exhibit a number of phenotypic differences. Common sunflower is characterised by many branches along its entire stem, each with numerous small heads and relatively small achenes. When disturbed, mature heads release their achenes, or "shatter". In contrast, cultivated sunflower is characterized by an unbranched stem, topped by a single large head, achenes, which are relatively large, are retained in the head until harvest. Domestication of plants from their wild progenitors has led to the production of a wide variety of crops that share a number of traits. Examples from the major cereals are larger grains, increased inflorescence size, more vigorous growth and loss of genes for shattering. The transition from small seeded plants with natural seed dispersal to larger seeded plants that retain their seeds until harvest, applies to all seed crops (Burke *et al.* 2002). Studies to understand the link between the phenotypic changes and the genes that are responsible are of paramount importance for the agriculture.

Harter *et al*. (2004) considered domesticated sunflower (*H. annuus*) to have had a single origin of domestication, thought to have arisen just once in the east-central United States. According to Wills and Burke (2006), Heiser (1985) discussed the possibility of an additional origin of domestication, perhaps in Mexico. This possibility has been raised by Lentz *et al*. (2001) after their discovery of carbonized achenes of sunflower in southern Mexico, which is beyond the current range of wild sunflower. Tang and Knapp (2003) examined the genetic diversity in sunflower as a whole using a set of nuclear simple sequence repeats. Their conclusion was that," the single ancestor hypothesis…seems improbable".

The domestication of crop plants is usually accompanied by a genome wide loss of genetic diversity (Tanksley and McCouch 1997). Together with domestication comes the transition to self-fertilization that can further reduce the levels of genetic diversity (Nordborg, 2000). Based on data from the major cereal crops, it appears that genome wide reductions in diversity are in the order of 30-40% (Buckler *et al*. 2001). Domestication can have a major impact on the organization of genetic diversity within the genome and therefore an increase in linkage disequilibrium (LD, the non-random association of alleles at different sites) throughout the genome (Liu *et al*. 2006). This is a recognized problem in sunflower, with a strong erosion of genetic diversity as breeding progresses because breeders tend to use the same elite germplasm in pursuit of similar breeding goals.

The diversity of the wild species is a valuable source of genes to introgress into the cultivated crop. Wild species of sunflower have a high level of genetic diversity as a consequence of their adaptation to the wide range of environments. Wild species harbour significant variability in a number of traits such as disease and pest resistance, quality of seeds and composition of compounds in seeds. Through interspecies crosses, breeders have transferred traits such as higher oil content, cytoplasmic male sterility (hybrid production), and insect and disease resistance to the cultivated sunflower. However, there are barriers preventing easy access to the genetic potential of the wild species. These include difficult cross ability, embryonic and post embryonic inter-specific and inter-generic incompatibility, and sterility in the F1 hybrid progeny (Encheva *et al.* 2003).

Incompatibility is typically overcome by a number of techniques, such as embryo rescue, ovular culture, somatic hybridization and callus culture that allow for the creation of a large number of inter-specific hybrids. Recent investigations have started looking at the possibilities of direct organogenesis. Hybrids of the wild species and cultivated sunflower have shown to have high regeneration potential (Yordanov *et al*. 2005). Yordanov and co-workers (2005) demonstrated the possibility to use a dendrogram as a methodology for early estimation of advantageous genotypes in plant selection for high regeneration potential. The use of biotechnology to move genes from other species into sunflower could speed up these interspecies crosses and recovery of genes tremendously. Molecular markers are also being used to explain partial hybridization in wide crosses between cultivated and perennial species of *Helianthus* (Faure *et al.* 2002).

1.2 Genetics of sunflower

Sunflower is a diploid (x=17) annual, with a basic chromosome number of 17 (subtribe Helianthea, subfamily Asteroidea, family Compositeae). Diploid, tetraploid and hexaploid species are known. The majority of the species are perennial and a few are annual. Despite its economic value, the number of simply inherited genes identified in sunflower is relatively small.

Genetic distance estimation for plant registration and protection using molecular markers (Lombard *et al*., 2001) is becoming increasingly important for international seed companies. However, there is virtually no information published about South African sunflower germplasm and therefore, this study is of considerable importance to South African sunflower breeders. It is important to the plant breeding community, and to commercial seed companies to have access to an economical and efficient analytic system that can offer an efficient and affordable system to perform variety verification (Mitchell *et al*., 1997; Senior *et al*., 1998).

According to Zhang *et al*. (2005), sunflower is strongly affected by the environment and the season, and most hybrids produce strong G x E interactions; the phenotype of the same hybrid may vary greatly according to location and the season. These factors make the implementation of distinctness, uniformity and stability using phenotyping a very difficult task. This has serious implications for seed companies, given that phenotypic traits are the defined characters used for registration and plant protection by UPOV (the Union Internationale pour la Protection des Obtentions Vegetales). For protection of Plant Breeders' Rights (PBR), parent inbred lines must be categorized in terms of distinctness, uniformity, and stability (DUS), using phenotypic trait descriptions. Due to rapid advancement in molecular techniques, the use of molecular markers in DUS testing as a complement to, or replacement of, morphological observations became the subject of great interest in scientific studies, and consequently a topic for discussion within UPOV. "Their integration into DUS testing protocols still depends upon resolving of several important issues. At this point in time, all DUS testing is still based on phenotypic evaluation of the plants", (Gunjaca *et al.* 2008).

The uniqueness of this study is the outright comparison of the DNA versus Protein markers used for genetic diversity study, through the use of SSR and Ultra-thin layer Iso-electric focusing (UTLIEF). Table 1 list some pros and cons of DNA versus Protein analyses for genetic diversity studies.

Table 1. Pros and Cons of using DNA versus Protein for the use of genetic diversity analysis, based on testing 96 samples for one data point

	DNA (SSR)	**Protein (UTLIEF)**
Cost	R 3 033.60	R 483.44
Time	4 hours	2 hours 35 min
Optimization	Once-off per marker	Continuous per different seeds sizes
Expression	Simple	Complex
Traits	Monogenetic only	Monogeneic or Polygenetic

1.3 Genetic analysis of sunflower using proteins for molecular markers

The value of molecular markers in sunflower genetic analysis has been demonstrated by several researchers. Isozymes have been used to assess genetic variation in both domesticated and wild sunflower populations (Cronn *et al*., 1997; Carrera *et al*., 2002), as well as to establish phylogenetic relationships and speciation mechanisms within the genus *Helianthus* (Reisberg *et al*., 1998). They have also been used to identify inter-specific hybrids (Carrera *et al*., 1996). Total protein fragment analysis has been used for phylogenetic studies in Russia in the last three years. Aksyonov (2005) used helianthin, a major seed protein, to establish the specificity of protein markers in sunflower and used albumin markers to define the genetic purity of sunflower.

1.3.1 Electrophoresis

The main fields of application for electrophoresis are biological and biochemical research, protein chemistry, pharmacology, forensic medicine, clinical investigations, veterinary science, food control as well as molecular biology (Westermeier, 2005).

Several forms of electrophoresis have been used to estimate the genetic diversity among different plant species (Hammes *et al*. 1990; Nasr, *et al* 2006). They have been used to estimate genetic diversity for phylogenetic reconstruction (Kaga *et al*. 1996), plant breeding, determination of relationships between varieties, development of linkage maps, and identification of markers connected with the resistance genes against pests and diseases. However, Tommasini *et al* (2003) cautioned that there is a limit to the degree of polymorphism that can be detected by biochemical and morphological markers and further, that these markers might be influenced by the environment and the stage of plant development when the plant samples are taken. In contrast, molecular markers are numerous, and are not affected by the environment or the age of the plant.

The execution of total protein genetic purity analysis may be based on the use of very high resolution ultrathin layer iso-electric focusing (UTLIEF) gels for the separation of a crude protein extracts into their components.

Iso-electric focusing (IEF) is an electrophoretic method that is limited to molecules which can either be positively or negatively charged i.e. proteins, enzymes and peptides (amphoteric molecules). Molecules thus separate according to their iso-electric points (pI), in a stabilized pH gradient. The net charge of a protein is the sum of all negative and positive charges of the amino acid side chains.

The method involves casting a layer of support media (usually a polyacrylamide gel or agarose). This medium contains a mixture of carrier ampholytes (low molecular weight synthetic polyamino-polycarboxylic acids). When using a polyacrylamide gel, a low percentage gel (~4%) is used because this has a large pore size, which allows proteins to move freely under the applied electrical field without hindrance. When an electric field is applied across such a gel, the carrier ampholytes arrange themselves in order of increasing pI from the anode to the cathode. Each carrier ampholyte maintains a local pH corresponding to its pI and thus a uniform pH gradient is created across the gel. If a protein sample is applied to the surface of the gel, then it will diffuse into the gel, and migrate up or down the gel, under the influence of the applied electric field until it reaches the region of the charge gradient where the pH corresponds to its iso-electric point. At this pH, the protein will have no net charge and will therefore become stationary at this point. Should the protein diffuse slightly toward the anode from this point, it will gain a weak positive charge and migrate back towards the cathode, to its position of zero charge. Similarly diffusion toward the cathode results in a weak negative charge that will direct the protein back to the same position. The protein is therefore trapped or “focused” at the pH value where it has zero charge. Proteins are therefore separated according to their charge, and not size as with SDS gel electrophoresis. Note that in IEF, it is crucial to find the correct place in the pH gradient to apply the sample, since some substances are unstable at certain pH values.

One problem with the use of UTLIEF to analyse protein profiles of sunflower is that visual interference often occurs in the gels. In other words, protein bands are not sharp and discrete, but instead, they run into neighbouring protein bands. This is primarily due to fats and oils naturally contained in the seeds. These are co-extracted with the proteins of choice, helianthinins and albumins.

1.3.2 Protein analysis

During the sunflower breeding and selection process, it is essential that genetic purity is maintained. Genetic purity is important for seed companies that guarantee their customers that they are purchasing high yielding hybrids with stable genetics, and designated traits such as resistance to certain diseases. Traditionally genetic purity analysis has been performed using phenotypic evaluations (Aksyonov, 2005). Typically, this consists of the physical inspection of sunflower plants at various stages of development, the flowering stage being the most important stage to assess purity. Unfortunately, this method has inherent flaws, according to Aksyonov, 2005, "the morphological parameters are neither sufficiently conspicuous nor sufficiently stable". Morphological properties are also affected by the environment (Sammour, 1991), as discussed above.

Electrophoretic protein markers were believed to be independent of cultivar morphology and physiology (Sammour, 1991). The advantages of using electrophoresis to identify these markers for variety and species identification are:

a. The process is relatively rapid;
b. It is relatively cheap;
c. It eliminates the need to grow plants to maturity;
d. The protein markers are largely unaffected by the environment.

However, there are some disadvantages in that protein markers may be influenced by tissue specificity and plant developmental stage. These disadvantages can be overcome by using seed storage proteins.

There are typically two classes of plant storage proteins: seed storage proteins (SSPs) and vegetative storage proteins (VSPs) (Fujiwara *et al*. 2002). SSPs accumulate to high levels in seeds during the late stages of seed development. They are degraded during seed germination, releasing amino acids to be utilized as protein building blocks for developing seedlings. The

SSPs determine the total protein content of the seed and the quality of the seed for consumers (Shewry *et al*. 1995). SSPs account for about 50% of the total protein in mature cereal grains (Shewry *et al*. 2005). SSP genes are classic targets for plant molecular biology. Their high expression in seed allowed for the development of techniques to detect of gene transcripts, and the development of cDNA cloning during the late 1970's and early 1980's (Fujiwara *et al*. 2002).

The detailed study of SSPs dates from the turn of the century, when Osborne (1924) classified them into groups on the basis of their extractability and solubility in water (albumins), dilute saline solutions(globulins), alcohol/water mixtures (prolamins), and dilute acids or alkalis (glutelins). The major seed storage proteins include the albumins, globulins and prolamins, according to "Osborne fractionation". A classification system used more recently places seed proteins into three groups: storage, structural and metabolic proteins.

Seed proteins were placed into two basic categories by Mandal *et al*. (2000), namely, housekeeping proteins and storage proteins. The housekeeping proteins are responsible for maintaining normal cell metabolism and this group of proteins can be further subdivided into storage, structural and biologically active proteins. Note that most physiologically active proteins are included in this group, i.e., lectins, enzymes and enzyme inhibitors. The SSPs are non-enzymatic and provide the amino acids required during germination and the establishment of new plants.

Storage globulins are contained in the embryo and outer aleurone layer of the endosperm. In maize these have been studied in some detail by Wallace and Kriz (1991). In sunflower there is an 11S globulin (helianthinin) that is a salt soluble protein, and is one of the major storage proteins of sunflower (Anisimova *et al*. 2004). The major endosperm storage proteins of all cereal grains are the prolamin storage proteins. All individual prolamin polypeptides are alcohol-soluble in the reduced state and vary greatly in molecular mass, from about 10 000 to almost 100 000. Prolamin has an evolutionary and structural relationship to the 2S albumin storage protein of sunflower, which is water-soluble.

Helianthinin is an oligomeric protein with a molecular mass (M_r) of approximately 305 000, consisting of six spherical subunits. This protein is characterised by the presence of several types of subunits and polypeptides, each with a different charge and M_r. The 2S albumins consist of a heterogeneous mixture of one-chain polypeptides with a M_r of about 10 000 – 18 000. (Anisimova *et al.* 2004).

The proteins of choice for molecular analysis of sunflower are the helianthinins and albumins. These are used as molecular markers to distinguish between sunflower cultivars, to check species identification, to assist biosystematic analysis and to study phylogenetic relationships of the species (Sammour, 1991).

An inbred protein marker can be described as a protein or proteins expressed in the hybrid progeny that was inherited from, and mono-morphic in, the inbred male of the hybrid, but polymorphic and absent in the inbred female of the hybrid; thus the presence of a self pollinated female will be clearly visible in the hybrid protein electro-phenogram. The analysis of protein markers allows for the reliable identification of homozygotes (lines) and heterozygotes (hybrids) in sunflower (Aksyonov, 2005).

The first step to hybrid sunflower production is the purification of the inbred lines involved in the hybrid crosses. Determination of this purity is therefore a key task for a seed company. Ultra-thin layer iso-electric focusing (UTLIEF) for the purpose of genetic purity analysis is currently the method of choice of some seed producing companies because this is a high throughput method that is cost effective, and which rapidly improves the genetic quality of the seed produced (van Oers and Tamboer, 2006).

1.4 Genetic analysis of sunflower using DNA for molecular markers

A growing number of genetic diversity studies have explored the use of nucleotide polymorphism data. These include studies on *Arabidopsis* (e.g., Savolainen *et al*. 2000; Aguade´ 2001; Nordborg *et al*. 2002; Wright *et al*. 2003; Ramos-Onsins *et al*. 2004), several major crops (e.g., White and Doebley 1999; Tenaillon *et al*. 2002; Garris *et al*. 2003; Zhu *et al*. 2003; Hamblin *et al*. 2004), and a handful of other taxa (e.g., Garcı´a-Gil *et al*. 2003; Kado *et al*. 2003; Brown *et al*. 2004; Ingvarsson 2005). Even though there are some similarities in these studies (e.g., a tendency toward reduced levels of polymorphism as a result of inbreeding) it is clear that the information gained from the study of any one system does not necessarily apply to another, even if they share similar mating systems, demographic histories, etc.

1.4.1 RFLP

During the last decade four restriction fragment length polymorphism (RFLP) linkage maps of cultivated sunflower have been published (Gentzbittel *et al*. 1999; Gedil *et al*. 2001). Two of the RFLP maps have been used as tools for mapping phenotypic and quantitative trait loci (Leon *et al*. 2003; Perez-Vich *et al*. 2002; Rachid Al-Chaarani *et al*. 2002). The widespread use of RFLP markers and maps in sunflower has been restricted by a lack of public RFLP probes and the low-throughput nature of RFLP markers (Yu *et al*. 2003).

1.4.2 RAPD

Concurrently, genetic-diversity and co-ancestry analyses have been carried out using random amplified polymorphic DNA (RAPD) analysis (Arias *et al*. 1995). RAPDs have primarily been used for tagging phenotypic loci in sunflower; e.g., resistance genes against rust (*Puccinia helianthi* Schw) and *Orobanche cumana* Wallr. (Lawson *et al*. 1998; Lu *et al*. 2000).

1.4.3 AFLP

The AFLP technique (amplified fragment length polymorphism) is considered an efficient marker system due to its high multiple applicability; e.g., for genetic mapping, DNA fingerprinting and diversity analysis (Kusterer *et al*. 2004). Cheres *et al* (1998) showed that AFLP is a powerful tool for the DNA fingerprinting of sunflower. AFLP has been used successfully in the establishment of genetic maps for several crop species, such as rice, maize and recently sunflower (Rachid Al Chaarani *et al*., 2002).

Although RAPD and AFLP markers have a multitude of uses, both are dominant, multi-copy, and often non-specific. As such, they are unsatisfactory for establishing a genome-wide framework of DNA markers for anchoring and cross referencing genetic linkage maps. The biggest negative to the use of AFLP in the commercial sector is the limited licence availability for commercial research

1.4.4 SSR

SSRs (simple sequence repeats), also called microsatellites, are widely used as molecular markers. They have become one of the principle classes of DNA markers used for DNA fingerprinting, genetic mapping, and molecular breeding in crop plants. SSR markers are preferred for several reasons:

a. SSRs are mostly multi-allelic and highly polymorphic (Jeffreys *et al*. 1994). SSR repeat length variants (alleles) are produced by DNA replication slippage and unequal crossing over between sister chromatids;
b. SSR markers can be genotyped rapidly and easily, using a variety of platforms for DNA fragment analysis, some of which are semi-automated (Cregan *et al*. 1999);

c. Details of SSR markers can be electronically dispersed and shared among laboratories;

d. SSR markers can be multiplexed by the length of the amplicon using virtually any electrophoretic system. When analysed using semi-automated, multicolour, genotyping systems, SSR markers can be doubled or tripled depending on the number of fluorophores supported by the system.

e. A large percentage of SSR markers, depending on the complexity of the host genome, amplify a single orthologous locus across genotypes.

1.4.5 Multiplex PCR

Multiplex PCR (Chamberlain *et al*. 1988) is a variation of the PCR technique used for applications where it is advantageous to amplify two or more loci simultaneously in the same reaction. In so doing, it can increase the amount of information generated per assay, and to reduce the costs of consumables and labour (Henegariu *et al*. 1997). This technique usually requires extensive optimization. The widespread use of multiplex PCR for SSR genotyping in crop plants has been limited by several factors. Firstly, PCR multiplexes have only been developed for a limited number of SSR on a few crops (Liu *et al*. 2000; Gethi *et al*. 2002). Secondly, the number of polymorphic SSR marker loci required for molecular breeding applications is often more that the number used in the multiplex PCR reactions. Thirdly, some SSR primers and primer combinations are recalcitrant to multiplex PCR procedures.

PCR-multiplexing is ideal for genotyping where common sets of SSR marker loci are required for repetitive DNA fingerprinting of new inbred lines and for rapid generation of inbred line identities. The protocols for multiplex PCR reactions, and the role of various ingredients in the multiplex PCR, have been described by several research groups (Henegariu *et al*. 1997; Zhang *et al*. 2003). The largest obstacles facing multiplex PCR are undesirable primer-primer interactions, and non-specific amplification (Elnifro *et al*. 2000). Another obstacle arises with the use of a tailed forward primer and a standard length reverse primer when the M13-tailed primer method is used because this can promote the amplification of non-specific products. Therefore, the PCR conditions required for amplification using the M13- tailed primer method are often different to those that are optimal for amplification using standard length primers.

1.4.6 Tailed PCR

The M13-tailed primer method (Oetting *et al*. 1995) is widely used for assays of SSRs, in order to reduce the cost of fluorescent primer labelling, which could be as much as five to ten times more expensive than the synthesis of an unlabeled primer. The M13-tailed primer method is a three primer strategy. Initially, a PCR is performed using a forward primer with a nucleotide extension at its 5'-end, identical to the sequence of an M13 sequencing primer (5'-CACGACGTTGTAAAACGAC-3'), a standard length reverse primer and a fluorescently labelled M13 primer. During the PCR, the SSR product is fluorescently labelled, following participation of the M13 primer after the first few cycles of amplification. Thus, instead of synthesizing one specific fluorescently labelled primer for each SSR marker, the labelled M13 primer is the sole source of label. As such, it can be used with any primer that contains the same sequence tail, and generates a labelled amplified DNA fragment.

Within a single amplification reaction, the PCR amplification occurs in two stages:

a. Amplicon 1 is produced using only the tailed forward and the 3' reverse primer. The extension of the forward primer yields a product that contains the "tail sequence". Thus when this template anneals with the reverse primer and extends, a product containing the complement of the tail sequence is produced (Amplicon 2).
b. The final step is the production of amplicon 3 by using the labelled M13 primer and Amplicon 2 as template. The fluorescent reporter is incorporated into the product during polymerization and a fluorescent signal is emitted. The DNA sequencer will only detect the labelled Amplicon 3. Figure 1 explains the protocol diagrammatically.

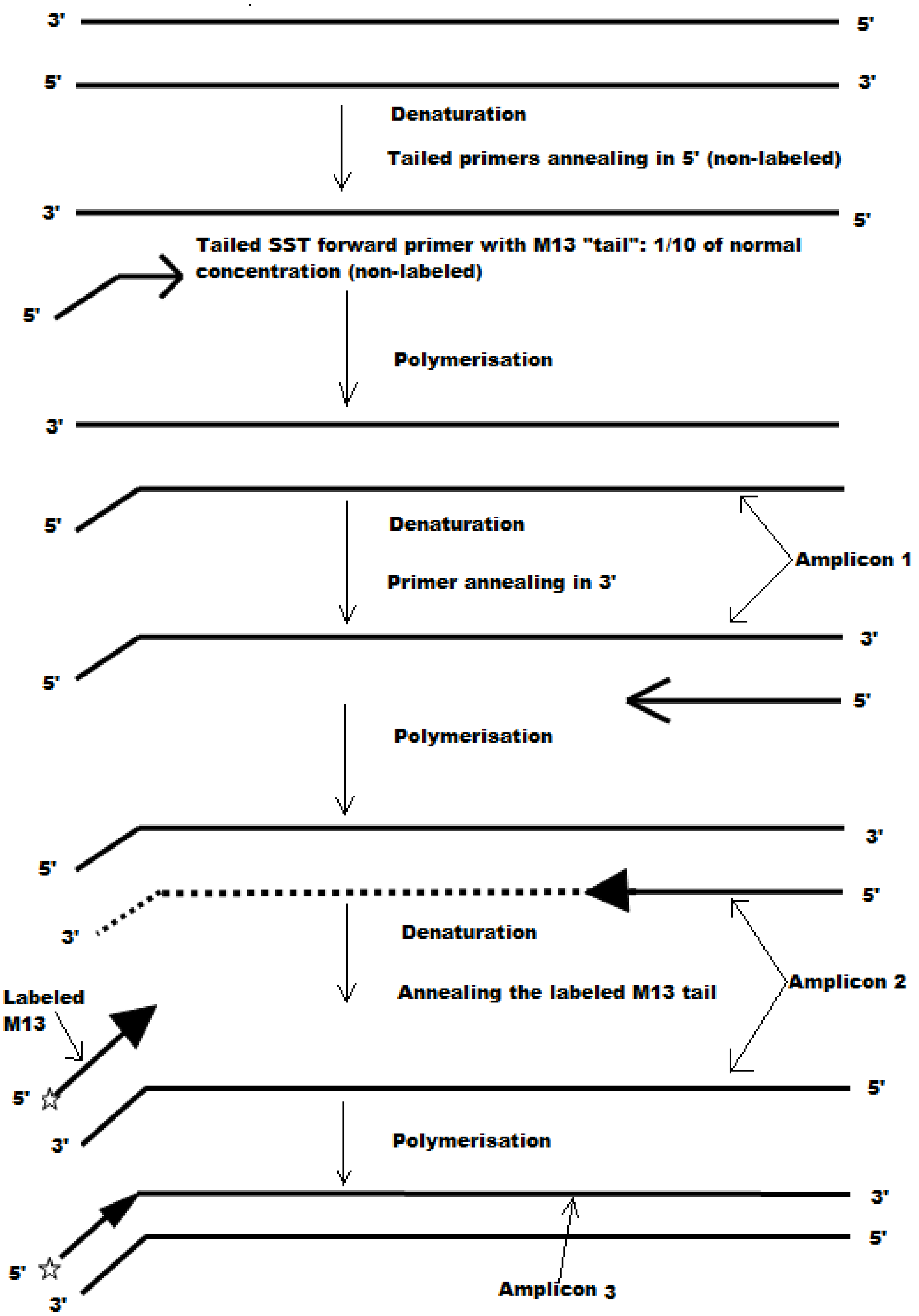

Figure 1. The M13-tailed primer method of PCR (Zhang *et al.* 2003)

Use of fluorescence-labelled microsatellite markers for genotyping on automated sequencers has many advantages over older techniques of SSR analysis that use auto radiographic or silver-stained detection techniques.

a. A large increase in throughput is made possible by the multiplexing of many PCR products into a single lane.
b. There is a significant increase in the accuracy of allele sizing, achieved by the use of internal size standards in each lane and of automated allele-calling algorithms.
c. This approach is much faster than conventional gel systems.
d. Automation of the process increases the speed and accuracy of data collection and processing.
e. The high sensitivity of detection also reduces the necessary volume (and therefore the cost) of the PCR reaction and allows detection of loci that are difficult to amplify.

Carrano *et al*. (1989) first reported on the use of fluorescence-based semi-automated analysis of marker panels. This method was adapted and improved upon for microsatellite analysis by Ziegle *et al*. (1992). Semi-automated methods of SSR genotyping, conducted by centralized laboratories, are rapidly replacing manual systems in plant breeding and genetics research. These methods facilitate the efficient application of microsatellite markers for high-throughput mapping (Tang *et al*. 2002; Zhang *et al*. 2005), pedigree analysis (Lexer *et al*. 1999), fingerprinting of accessions (Carrano *et al*. 1989), and assaying for genetic diversity (Macaulay *et al*. 2001; Zhang *et al*. 2005). The technology has multiple applications for the seed industry: it can improve the efficiency of managing a germplasm collection, help deliver purity-proven seed stocks to growers, and provide the basis for PBR protection (Mitchell *et al*. 1997).

1.5 Genetic distance

Genetic distance estimations using molecular markers are becoming increasingly important for international seed companies for plant registration and PBR protection.

According to Yu *et al* (2002; 2003), the development of 1089 SSR markers for cultivated sunflower eliminated a long-standing bottleneck caused by the scarcity of single-copy DNA markers in the public domain. Tang *et al*. (2002) constructed the first genetic linkage map of sunflower on the basis of SSR markers and the first dense public genetic linkage map on the basis of single or low-copy DNA markers.

Understanding the genetic diversity of parental lines is crucial to the success of plant breeding programmes, in particular when the objective is the production of hybrid seed. This information gives a breeder clarity about heterotic groups and therefore crosses of parental lines with the most potential to maximise heterosis.

A complication was identified by Burstin *et al*. (1994) who commented, "pedigree information provides a global estimate of the expected genetic relatedness among lines, but relies on the assumption of the absence of gametic and zygotic selection, which is often not the case".

An increasing number of molecular markers have been correlated with morphological and biochemical data, to assess genetic diversity among parental lines. Data sets have been compiled to reflecting genetic diversity based on morphology (Bar-Hen *et al*. 1995), isozymes (Hamrick and Godt, 1997) and storage protein profiles (Smith *et al*. 1987). In recent years the use of DNA markers has been proposed for "precise and reliable characterization and discrimination of genotypes" (Jaikishen, *et al*. 2004).

1.6 References

Aguade´, M., 2001. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes in *Arabidopsis thaliana*. Molecular Biological Evolution 18: 1–9.

Aksyonov, I.V. 2005.Use of albumin markers for defining genetic purity of sunflower parent lines and hybrids. Helia 28: 43-48.

Anisimova, I.N., Gavrilova, V.A., Loskutov, A.V., Rozhkova, V.T. and Tolmachev, V.V. 2004. Polymorphism and inheritance of seed storage protein in sunflower. Russian Journal of Genetics 40: 995-1002.

Arias, D.M. and Reiseberg, L.H. 1995. Genetic relationship among domesticated and wild sunflower (*Helianthus annuus*, Asteraceae). Economical Botany. 49: 239-248.

Bar-Hen, A., Charcosset, A., Bourgoin, M. and Cuiard, J. 1995. Relationships between genetic markers and morphological traits in a maize inbred lines collection. Euphytica 84: 145-154.

Burstin, J., de Vienne, D., Dubreuil, P. and Damerval, C. 1994. Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 inbred lines. Theoretical and Applied Genetics 89: 943-950.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H. and Neale, D.B. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. Procedures of the National Academy of Science USA 101: 15255–15260.

Buckler, E.S., Thornberry, J.F., and Kresovich, S. 2001. Molecular diversity, structure and domestication of grasses. Genetic Resources 77: 213-218.

Burke, J.M., Tang, S., Knapp, S.J. and Rieseberg, L.H. 2002. Genetic analysis of sunflower domestication. Genetics 161: 1257-1267.

Carrano, A.V., Lamerdin, J., Ashworth, L.K., Watkins, B., Branscomb, E., Slezak, T., Raff, M., De Jong, P.J., Keith, D., McBride, L., Meister, S. and Kronick, M. 1989. A high resolution, fluorescence based, semi-automated method for DNA fingerprinting. Genomics 4: 129–136.

Carrera, A., Poverene, M. and Rodriguez, R.H. 1996. Isozyme variability in *Helianthus argophyllus*. Its application in crosses with cultivated sunflower. Helia 19: 19-28.

Carrera, A.D., Pizarro, G., Poverene, M., Feingold, S., León, A.J. and Berry, S.T. 2002. Variability among inbred lines and RFLP mapping of sunflower isozymes. Genetics and Molecular Biology 25: 65-72.

Chamberlain, J.S., Gibss, R.A., Ranier, J.E., Nguyen, P.N. and Caskey, C.T. 1988. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acid Research 16: 11141-11156.

Chen, J., Hu, J., Vick, B.A. and Jan, C. C. 2006. Molecular mapping of a nuclear male-sterility gene in sunflower (*Helianthus annuus* L.) using TRAP and SSR markers. Theoretical and Applied Genetics 113: 122-127.

Cheres, M.T. and Knapp, S. 1998. Ancestral origins and genetic diversity of cultivated sunflower: co-ancestry analysis of public germplasm. Crop Science 38: 1476-1482.

Cooke, R.J. 1984. The characterization and identification of crop cultivars by electrophoresis. Electrophoresis 5: 59-72.

Cregan, P.B., Jarvik, T., Bush, A.L., Shoemaker, R.C., Lark, K.G., Kahler, A.I., Kaya, N., Van Toai, T.T., Lohnes, D.G., Chung, J. and Specht, J.E. 1999. An integrated genetic linkage map of the soybean genome. Crop Science 39: 1464-1490.

Cronn, R., Brothers, M., Klier, K., Bretting, P.K. and Wendel, J.F. 1997. Allozyme variation in domesticated annual sunflower and its wild relatives. Theoretical and Applied Genetics 95: 532-545.

Elnifro, E.M., Ashshi, A.M., Cooper, R.J. and Klapper, P.E. 2000. Multiplex PCR: optimisation and application in diagnostic virology. Clinical Microbiology Reviews 13: 559-570.

Encheva, J., Christov, M., and Ivanov, P. 2003. Characterization of interspecific hybrids between cultivated sunflower *H. annuus* L. (cv. Albena) and wild species *Helianthus tuberosus.* Helia 26 Nr 39: 43-50.

Faure, N., Serieys, H., Cazaux, E., Kaan, F, and Berville, A. 2002. Partial hybrization in wide crosses between cultivated sunflower and the perennial *Helianthus* species *H. mollis* and *H. orgyalis*. Annals of Botany 89: 31-39.

Fujiwara, T., Nambara, E., Yamagishi, K., Goto, D.B. and Naito, S. 2002. Storage proteins. American Society of Plant Biology. The Arabidopsis Book. http://www.aspb.org/publications/arabidopsis

Garcı´a-Gil, M.L., Mikkonen, M. and Savolainen, O. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Molecular Ecology 12: 1195–1206.

Garris, A.J., McCouch ,S.R. and Kresovich, S. 2003 Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). Genetics 165: 759–769.

Gedil, M. A., Wye, C., Berry, S., Segers, B., Peleman, J., Jones, R., Leon, A., Slabaugh, M.B. and Knapp, S.J. 2001. An integrated restriction fragment length polymorphism-amplified fragment length polymorphism linkage map for cultivated sunflower. Genome 44: 213-221.

Gentzbittel, L., Mestries, E., Mouzeyar, S., Mazeyrat, F., Badaoui, S., Vear, F., Tourvieill de Labrouhe, D. and Nicolas, P. 1999. A composite map of expressed sequences and phenotypic traits of the sunflower (*Helianthus annuus* L.) genome. Theoretical and Applied Genetics 99: 218-234.

Gethi, J.G. Labate, J.A., Lamkey, K.R., Smith, M.E. and Kresovich, S. 2002. SSR variation in important U.S. maize inbred lines. Crop Science 42: 951-957.

Gunjaca, J., Buhinicek, I., Jukic, M., Sarcevic. H., Vragolovic, A., Kozic, Z., Jambrovic, A. and Pejic, I. 2008. Discriminating maize inbred lines using molecular and DUS data. Euphytica 161: 165-172.

Hamblin, M.T., Mitchell, S.E., White, G.M., Gallego, J., Kukatla, R. 2004. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolour*. Genetics 167: 471–483.

Hammes, B.D. and Richwood, M. 1990. Gel Electrophoresis of Proteins, a Practical Approach. Oxford University Press, UK.

Hamrick, J.L. and Godt, M.J.W. 1997. Allozyme diversity in cultivated crops. Crop Science 37: 26-30.

Harter, A.V., Gardner, K.A., Falush, D., Lentz, D.L., Bye, R.A. and Rieseberg, L.H. 2004. Origin of extant domesticated sunflowers in eastern North America. Nature 430: 201-205.

Henegariu, O., Heerema, N.A., Dlouhy, S.R., Vance, G.H. and Vogt, P.H.1997. Multiplex PCR: Critical parameters and step-by-step protocol. Biotechniques 23: 504-511.

Hvarleva, T., Bakalova, A., Chepinski, I., Hristova-Cherbadji, M., Hristov, M. and Atanasov, A. 2007. Characterization of Bulgarian sunflower cultivars and inbred lines with microsatellite markers. Biotechnology and Biotechnology Equipment 21: 408-412.

Ingvarsson, P.K. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus termula* L., Salicaceae). Genetics 169: 945–953.

Jaikishen, I., Ramesha, M.S., Rajendrakumar, P. Rao, K.S., Neeraja, C.N. Balachandran, S.M., Viraktamath, B.C., Sujatha, K. and Sundaram, R.M. Characterization of genetic diversity in hybrid rice parental lines using EST-derived and non-EST SSR markers. Rice Genetics Newsletter 23: 24-28.

Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D. G. Neil, D.L. and Armour, J.A.L. 1994. Complex gene conversion events in germline mutations at human minisatellites. Nature Genetics 6: 136-145.

Kado, T., Yoshimaru, H., Tsumura, Y. and Tachida, H. 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae *sensu lato*). Genetics 164: 1547–1599.

Kaga. A., Tomooka, N., Egava, Y., Hosaka, K. and Kamijima, O. 1996. Species relationship in the subgenus *Ceratotropis* (genus *Vigna*) as revealed by RAPD analysis. Euphytica 88: 17-24.

Kusterer, B. Rozynek, B. Brahm, L., Prufe, M., Tzigos, S. Horn, R. and Friedt, W. 2004. Construction of a genetic map and localization of major traits in sunflower. Helia 27 Nr 40: 15-24.

Lawson, W.R., Goulter, R.J., Henry, R.J., Kong, G.A. and Kochman, J.K. 1998. Marker-assisted selection for two rust resistance genes in sunflower. Molecular Breeding 4: 227-234.

Le clercq, P. 1969. Une sterilite male cytoplasmique chez le tournesol. Annual Amelior Plant 19: 99-106.

Lentz, D.I., Pohl, M.E.D., Pope, K.O., and Wyatt, A.R. 2001. Prehistoric sunflower (*Helianthus annuus* L.) domestication in Mexico. Economic Botany 55: 370-376.

Leon, A.J., Andrade, F.H. and Lee, M. 2003. Genetic analysis of seed oil concentration across generations and environments in sunflower (*Helianthus annuus* L.). Crop Science 43: 135-140.

Liu, S. Cantrell, R.G., McCarty, J.C. and Stewart, J.M. 2000. Simple Sequence Repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Science 40: 1459-1469.

Liu, A. and Burke, J.M. 2006. Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics 173: 321-330.

Lombard, L., Dubreuil, P., Dillman, C. and Baril, C. 2001. Genetic distance estimator based on molecular data of plant registration and protection: a review. Acta Horticulturae 546: 55-63.

Lu, Y.H., Melero-Vara, J.M., Garcia-Tejada, J.A. and Blanchard. 2000. Development of SCAR markers linked to the gene *Or5* conferring rust resistance to broomrape (*Orobanche cumana* Wallr.) in sunflower. Theoretical and Applied Genetics 100: 625-632.

Macaulay, M., Ramsay, L., Powell, W. and Waugh, R. 2001. A representative, highly informative genotyping set of barley SSRs. Theoretical and Applied Genetics 102: 801–809.

Mandal, S. and Mandal, R.K. 2000. Seed storage proteins and approaches for the improvement of their nutritional quality by genetic engineering. Current Science 79: 576-589.

Mitchell, S.E., Kresovich, S., Jester, C.A., Hernandez, C.J. and Szewe-McFadden, A.K. 1997. Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Science 37: 617-624.

Nasr. N., Khayami, M., Hedari, R. and Jamei, R. 2006. Genetic diversity among selected varieties of *Brassica napus* (Crucifereae) based on the biochemical composition of seeds. Journal of Science (University of Tehran) 32: 37-40.

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

Nordborg, M., Borevitz, J.O., Bergelson, J., Berry, C.C., Chory, J. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. National Genetics 30: 190–193.

Oetting, W.S., Lee, H.K., Flanders, D.J., Wiesner, G.L., Sellers, T.A. and King, R.A. 1995.Linkage analysis with multiplexed short tandem repeat polymorphisms using infra-red fluorescence and M13 tailed primers. Genomics 30: 450-458.

Osborne, T.B. 1924. The Vegetable Proteins. Longmans, Green, London.

Paniego, N., Echaide, M., Muňoz, M., Fernăndez, L., Torales, S., Faccio, P., Fuxan, I., Carrera, M., Zandomeni, R., Suărez, E.Y. and Hopp, H.E. 2002. Microsatellite isolation and characterization in sunflower (*Helianthus annuus* L.). Genome 4: 34-43.

Perez-Vich, B. Fernandez-Martinez, J.M., Grondona, M., Knapp, S.J. and Berry, S.T. 2002. Stearoyl-ACP and oleoyl-PC desaturase genes co-segregate with quantitative trait loci underlying high stearic and high oleic acid mutant phenotypes in sunflower. Theoretical and Applied Genetics 104: 338-349.

Putnam, D.H., Oplinger, E.S., Hicks, D.R., Durgan, B.R., Noetzel, D.M., Meronuck, R.A., Doll, J.D. and Schulte, E.E. 1990. Alternative Field Crops Manual: Sunflower.
http://www.hort.purdue.edu/newcrop/afcm/sunflower.html
Downloaded 10.12.2008

Rachid Al-Chaarani, G., Roustaee, A., Gentzbittel, L., Modrani, L., Barrault, G., Dechamp-Guillaume, G. and Sarrafi, A. 2002. A QTL analysis of sunflower partial resistance to downy mildew (*Plasmopara halstedii*) and black stem (*Phoma macdonaldii*) by the use of recombinant inbred lines (RILs). Theoretical and Applied Genetics 104: 490-496.

Ramos-Onsins, S.E., Stranger, B. E., Mitchell-Olds, T. and Aguade´, M. 2004 Multilocus analysis of variation and specialization in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics 166: 373–388.

Reiseberg, L., Baird, S.J.E. and Desrochers, A. M. 1998. Patterns of mating in wild sunflower hybrid zones. Evolution 52: 713-726.

Sammour, R.H. 1991. Using electrophoretic techniques in varietal identification, biosystematic analysis, phylogenetic relations and genetic resources management. Journal of Islamic Academy of Sciences 4: 221-226.

Savolainen, O., Langley, C.H., Lazzaro, B.P. and Fre´ville, H. 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Molecular Biological Evolution 17: 645–655.

Senior, M.L., Murohy, M.M., Goodman, M.M., Stuber, C.W. 1998. Utility of SSRs for determining genetic similarities and relationships in maize using agarose gel systems. Crop Science 38: 1088-1098.

Shewry, P.R., Napier, J.A. and Tatham, A.S. 1995. Seed storage proteins: structure and biosynthesis. The Plant Cell 7: 945-956.

Shewry, P.R. and Halford, N.G. 2005. Cereal seed storage proteins: structure properties and role in grain utilization. Journal of Experimental Botany 53: 947-958.

Smith, J.S.C., Paszkiewicks, S., Smith, O.S. and Schaeffer, J. 1987. Electrophoretic, chromatographic and genetic techniques for identifying associations and measuring genetic diversity among corn hybrids. In Proceedings 42nd Annual Corn Sorghum Research Conference. Chicago, IL. American Seed Trade Association., Washington, DC. 187-203.

Tang, S., Yu, J.K., Slabaugh, M.B., Shintani, D.K. and Knapp, S.J. 2002. Simple sequence repeat map of the sunflower genome. Theoretical and Applied Genetics 105: 1124-1136.

Tang, S. and Knapp, S.J. 2003. Microsatellites uncover extraordinary diversity in native American land races and wild populations of cultivated sunflower. Theoretical and Applied Genetics 16: 990-1003.

Tanksley, S.D., and McCough, S.R., 1997. Seed banks and molecular maps; unlocking genetic potential from the wild. Science 277: 1063-1066.

Tenaillon, M.I., Sawkins, M.C., Anderson, L.K., Stack, S. M. and Doebley, J. 2002. Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics 162: 1401–1413.

Tommasini, L., Batley, J., Arnold, G.M., Cooke, R.J., Donini, P., Lee, D., Law, J.R., Lowe, C., Moule, C., Trick, M. and Edwards, K.J. 2003. The development of multiplex simple sequence repeats (SSR) markers to compliment distinctness, uniformity and stability testing of rape (*Brassica napus* L.) varieties. Theoretical and Applied Genetics 106: 1091-1101.

van Oers, C. M. and Tamboer, J. H. A. 2006. Discriminative power of ultra thin layer gel systems in acrylamide isoelectric focusing for verification of varieties with a very narrow genetic variation [unpublished handout]. Proteios BV, the Netherlands.

Wallace, N.H. and Kriz, A.L. 1991. Nucleotide sequence of a cDNA clone corresponding to the maize Globulin-2 gene. Plant Physiology 95: 973-975.

Westermeier, R. 2005. Electrophoresis in Practice, Fourth Edition. Wiley-VCH, Weinheim, Federal Republic of Germany.

White, S.E., and Doebley, J.F. 1999. The molecular evolution of terminal ear1, a regulatory gene in the genus *Zea*. Genetics 153: 1455–1462.

Wills, D.M., and Burke, J.M. 2006. Chloroplast DNA variation confirms a single origin of domesticated sunflower (*Helianthus annuus* L.) Journal of Heredity 97(4): 403-408.

Wright, S., Lauga, B. and Charlesworth, D. 2003. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Molecular Ecology 12: 1247–1263.

Yordanov, Y., Atanassov, I., E., Yordanova, E., Atanassov, A., Georgiev, S. and Christov, M. 2005a. Characterization of backcross lines of Helianthus *eggertii* Small x *Helianthus annuus* L. possessing different regeneration capacity by DNA and isozyme markers. Biotechnology and Biotechnology Equipment 19 (1): 57-62.

Yordanov, Y., Hristov, E., Yordanova, E., Atannassov, I. and Georgiev, S. 2005. Using DNA and isozyme markers to study genetic relationship among high regenerative interspecific hybrids of *Helianthus eggertii* Small x *Helianthus annuus* L. General and Applied Genetics 19: 27-32.

Yu, J., Mangor, J., Thompson, L., Edwards, K.J., Slabaugh, M.B. and Knapp, S.J. 2002. Allelic diversity of simple sequence repeats among elite inbred lines of cultivated sunflower. Genome 45: 652-660.

Yu, J., Tang, S., Slabaugh, M.B., Heesacker, A., Cole, G., Herring, M., Soper, J., Han, F., Chu, W., Webb, D.M., Thompson, L., Edwards, K.J., Berry, S., Leon, A.J., Grondona, M., Olungu, C., Maes, N. and Knapp, S.J. 2003. Towards a saturated molecular genetic linkage map for cultivated sunflower. Crop Science 43: 367-387.

Zhang, L.S., Bacquet, V., Li, S.H. and Zhang, D. 2003. Optimization of multiplex PCR and multiplex gel electrophoresis in sunflower SSR analysis using infrared fluorescence and tailed primers. Acta Botanica Sinica 45 (11): 1312-1318.

Zhang, L.S., Le Clerc, V., Li, S. and Zhang, D. 2005. Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. Canadian Journal of Botany 83: 66-72.

Zhu, Y.L., Song , Q.J., Hyten, D.L., Van Tassell, C.P. and Matukumalli, L.K. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163: 1123– 134.

Ziegle, J.S., Su, Y., Corcoran, K.P., Nie, L., Mayrand, P.E., Hoff, L.B., McBide, L.J., Kronick, M.N. and Diehl, S.R. 1992. Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics 14: 1026–1031.

CHAPTER 2: Variety identification and genetic diversity of proprietary inbred lines of sunflower, determined by mapped SSR markers

2.1 Abstract

The oilseed sunflower (*Helianthus annuus* Linnaeus) gene pool is the product of multiple breeding and domestication bottlenecks. Early genetic studies have led to the hypothesis of a single point of domestication. The objectives of this study were (i) to assess the level of genetic diversity in elite maintainer line (B line) and fertility-restoring (R) sunflower lines in a proprietary breeding programme; and (ii) to compare the classification of germplasm on the basis of estimates of genetic similarities obtained by means of microsatellite (SSR) markers. A set of 73 SSR markers with known mapped positions were utilized to determine the genetic similarity in a group of B and R inbred lines of sunflower. Cluster analysis of genetic similarity revealed an excellent correlation with the breeding background and source information obtained from breeders on all inbred lines used in this study. Cluster analysis gave a clear differentiation between B and R-lines, showing clearly defined heterotic groups of the proprietary set of inbred lines.

2.2 Introduction

Genetic distance estimation for plant registration and protection using molecular markers is becoming increasingly important for international seed companies. There is virtually no information published about proprietary African sunflower material and this study is of high importance to breeders in the industry. It is important in the scientific and commercial environment to have an economical and efficient analysis system to perform variety verification (Mitchell *et al.*, 1997; Senior *et al.*, 1998), and fingerprinting on large study populations.

Cultivated sunflower cultivars are produced as hybrids, obtained by crossing a male-sterile, female inbred line (A line) with a restorer male line (R line). The sterility of the A line is maintained by crossing it with its isogenic fertile line (B line). For legal plant protection according to UPOV (Union Internationale pour la Protection des Obtentions Vegetales), the parent inbred lines must demonstrate distinctness, uniformity, and stability using phenotypic trait descriptions. The genetic base for sunflower breeders is slowly being reduced, due to the frequent use of the same genetic resources for common breeding objectives (i.e. seed yield and resistance). According to Zhang *et al.*, (2005), "sunflower is a plant very sensitive to interactions among genotype, location, and year; the phenotype of the same plant may vary greatly on the same plant material, according to location and the growing year". These factors make the use of phenotypic means of registration and plant protection of sunflower cultivars very difficult because demonstrating distinctness, uniformity and stability in sunflower is extremely challenging when based only on phenotypic data.

A relatively small, but growing number of studies, look at plants genotypes using nucleotide polymorphism data such as *Arabidopsis* (e.g., Savolainen *et al.*, 2000; Aguade, 2001;Nordborg *et al.*, 2002;Wright *et al.*, 2003; Ramos-Onsins *et al.*, 2004), in several major crops (e.g., White and Doebley, 1999; Tenaillon *et al.*, 2002; Garris *et al.*, 2003; Zhu *et al.*, 2003; Hamblin *et al.*, 2004), and a handful of other taxa (e.g., Garcı´a-Gil *et al.*, 2003; Kado *et al.*, 2003; Brown *et al.*, 2004; Ingvarsson, 2005). Even though there are some similarities in these studies (e.g., a tendency toward reduced levels of polymorphism), it is clear that the information gained from the study of any one system do not necessarily apply to another, even if they share similar mating systems, demographic histories, etc.

The importance of molecular markers in sunflower genetic analysis has been demonstrated by several studies. Isozymes have been used to assess genetic variation in both domesticated and wild sunflower populations (Cronn *et al.*, 1997; Carrera *et al.*, 2002), as well as to establish phylogenetic relationships and speciation mechanisms within the genus *Helianthus* (Reisberg *et al.*, 1998). They have also been used to identify interspecific hybrids (Carrera *et al.*, 1996). Total protein fragment analysis has been used to detect molecular markers for phylogenetic studies in Russia in the last three years. Aksyonov (2005) used helianthin, a major seed protein, to establish the specificity of protein markers in sunflower and used albumin markers to define the genetic purity of sunflower.

During the last decade four restriction fragment length polymorphism (RFLP) linkage maps of cultivated sunflower were published (Gentzbittel *et al.*, 1999; Gedil *et al.*, 2001). Simultaneously, genetic-diversity and co-ancestry analyses were carried out using random amplified polymorphic DNA (RAPD) (Arias *et al.*, 1995). RAPDs have primarily been used for tagging phenotypic loci in sunflower, for example, resistance genes to rust (*Puccinia helianthi* Schwein) and *Orobanche cumane* Wallroth. (Yu *et al.*, 2003). The AFLP technique (amplified fragment length polymorphism) is considered an efficient marker system due to its high multiple applicability, e.g., genetic mapping fingerprinting and diversity analysis. Hongtrakul *et al.* (1997) showed that AFLP can be a powerful tool for fingerprinting of sunflower. AFLP has been used successfully in the establishment of genetic maps in several crop species, including rice, maize and sunflower (Rachid Al Chaarani *et al.*, 2001). The biggest problem with the use of AFLPs in the commercial sector is the limited licence availability for commercial research. Even though RAPD and AFLP have a multitude of uses, both are dominant, multicopy, and are often non-specific in nature.

SSRs (simple sequence repeats), also called microsatellites, are widely used as molecular markers. SSRs are short sequence elements arranged in simple internal repeat structures (Paniego *et al.*, 2002) that are densely and randomly distributed throughout eukaryotic genomes. According to Hvarleva *et al.* (2007), SSRs are the most reliable markers for cultivar identification, genetic diversity evaluation and intellectual property rights protection. Because of their high rates of polymorphism, random distribution and co-dominant

Mendelian inheritance and high mutation rate, they constitute the molecular markers with the highest polymorphic information content (PIC). Microsatellites that are high in polymorphism have co abundance and high levels of distribution throughout plant genomes. SSRs have become one of the principle classes of DNA markers used for DNA fingerprinting, genetic mapping, and molecular breeding in crop plants (Morgate *et al.*, 1993). There are various reasons for the preferred use of SSR markers. Firstly, SSRs are mostly multi-allelic and highly polymorphic (Jeffreys *et al.* 1994). SSR repeat length variants (alleles) are produced by DNA replication slippage and unequal crossing over between sister chromatids. Secondly, SSR markers can be genotyped rapidly using a variety of platforms for DNA fragment analysis, some of which are semi-automated (Cregan *et al.*, 1999). Thirdly, the identity of SSR markers can be electronically dispersed and shared among laboratories. Fourthly, SSR markers can be multiplexed by the length of the amplicon using virtually any electrophoretic system. When analysed using semi-automated, multicolour, genotyping systems, SSR markers can be doubled or tripled depending on the number of fluorophores supported by the system. Fifthly, a large percentage of SSR markers, depending on the complexity of the host genome, amplify a single orthologous locus across genotypes.

According to Yu *et al* (2003) the development of 1089 SSR markers for cultivated sunflower eliminated the long-standing bottleneck caused by the scarcity of single-copy DNA markers in the public domain (Yu *et al.*, 2002). Tang *et al.* (2002) constructed the first genetic linkage map of sunflower on the basis of SSR markers and the first dense public genetic linkage map on the basis of single or low-copy DNA markers.

This study describes the use of SSR marker systems for the investigation of allelic diversity of *Helianthus*, and the relatedness of a set of inbred lines.

2.3 Materials and Methods

2.3.1 Plant materials and isolation of DNA

Genomic DNA was isolated from 7 day old seedlings, grown under controlled conditions. Five individuals per germplasm accession were harvested. Approximately 400mg of young leaf tissue was put into a mortar and manually ground under liquid nitrogen. A 100mg of the frozen ground leaf material were weighed into an eppendorf vial and its DNA was extracted using a Sigma Nucleic Extraction kit, according to the supplier's specifications.

DNA was isolated from 33 inbred lines. The DNA concentration was determined using 0.7% TBE agarose. A working concentration of 10ng μl^{-1} was standardized on all extracted DNA. Among the material extracted were 20 male restorer lines and 13 female maintainer lines. Some of these lines had special relationships, e.g., the normal (TF152R) and the downy mildew resistant version (TF152RRM) of the same inbred line and the normal (TF152R) and high oleic acid version (TF152RHL) of the same inbred line.

2.3.2 Microsatellite genotyping

Microsatellite genotypes were produced for 33 elite inbred lines using 73 microsatellite markers selected from a public collection (Tang *et al.*, 2002; Yu *et al.*, 2002). SSR genotyping primers were synthesized by Inqaba Biotech SA, and the fluorescent tails were synthesized by Applied Biosystems, Johannesburg, South Africa.

SSR genotyping were performed using an ABI3130xl (Applied Biosystems, Johannesburg, South Africa) sequence analyzer. Genotypes were ascertained using MapMaker 3.1, from Applied Biosystems.

PCR reactions were performed using 12µl of a reaction mixture containing 1 x PCR buffer, 2.5mM Mg^{++}, 0.2µl each of dNTPs (Bioline), 1 unit of Taq polymerase (Bioline) and 5-10ng of genomic DNA. Primers were labelled with a fluorescent dye; using a tailed primer strategy (Zhang *et al.*, 2005). One tail, M13 (5'-CACGACGTTGTAAAACGAC-3'), was added to 5'-end of one of the SSR primers (forward primer) during primer synthesis. Three primers are required for the amplification of each SSR locus: one tailed forward primer (0.05µmol), one normal reverse primer (0.25µmol) and one labelled tail (0.2µmol) were used.

A "Touchdown" PCR was used to reduce spurious amplification. The initial denaturation step was performed at 94ºC for 2min, followed by 1 cycle at 94ºC for 30s, 63ºC for 30s and 72ºC for 45s. The annealing temperature was decreased by 1ºC per cycle in subsequent cycles until it reached a temperature of 57ºC. Products were subsequently amplified for 32 cycles at 94ºC for 30s, 57ºC for 30s, and 72ºC for 45s with a final extension for 20min.

Amplified loci were detected by laser scanning during electrophoresis, using an ABI 3130xl Sequencer (Applied Biosystems). Samples containing 1µl of the PCR products were mixed with 8.5µl loading buffer (formamide) and 0.5µl Liz-250 internal standard (ABI). Samples were denatured at 95ºC for 5min and cooled to 4ºC and loaded on the auto-sampler for auto injection and capillary electrophoresis. Band sizes were generated automatically in comparison with a standard sizing ladder included in every sample prior to electrophoresis, using Genescan® and Genotyper® computer software, from ABI. Band scoring was then checked manually.

2.3.3 Data collection and analysis

The amplification profile for each microsatellite was scored semi-automatically and evaluated. Ambiguous data were re-examined and scored manually. Bands with the same mobility were considered identical, receiving equal values. SSR markers were usually considered to reveal a single locus per primer combination. The presence of only one allele of a given microsatellite was considered a homozygous state of the allele, assuming the absence of null alleles.

The availability of marker data allows comparison of genotypes for these marker data. An overall analysis of the relatedness of all genotypes in the data set can be performed by calculating the genetic distance for each pair of genotypes. There are several measures for estimating the genetic distance based on the marker data. For this analysis two types of analysis were investigated: (1) the Jaccard distance that is the simple matching coefficient (the number of shared alleles as a proportion of all alleles); and (2) the Euclidean distance (the square root of the sum of all squared differences between alleles. The Euclidean distance is often used for quantitative data and is somewhat artificial for re-coded marker data.

Genetic distance was measured by evaluating the proportion of shared allele's per locus, polymorphic information content (PIC) and similarity values. The inbred lines were fingerprinted and therefore the selected inbreds were presumed to be homozygous for most loci. PIC estimated the probability of observing a polymorphism between two inbred lines, randomly drawn from the sample of 33.

A graphical representation the molecular marker data was obtained by using a programme called "GGT" (an acronym for Graphical Geno Types).(Ralph van Berloo., 2007). The data was imported into this programme making use of commonly used maker file types that contain certain marker information. GGT data files were derived from two sources of data: A locus file, containing marker names and raw marker scored and a (linkage) map file, specifying marker positions on a linkage map.

2.4 Results

Thirty three inbred lines were genotyped using 73 mapped microsatellite markers. The markers are dispersed throughout the sunflower genome. The selected microsatellite markers each amplified a single locus across the 33 germplasm accessions. The SSR markers were screened for polymorphisms among the 33 inbred lines to estimate allele-length ranges, assess genotyping qualities, and to identify SSR markers for testing in PCR multiplexes. Table 1 shows the list of 73 markers used for this study.

Table 1. 73 Sunflower simple sequence repeat (SSR) markers, showing mapped position, expected allele lengths, linkage groups and polymorphic information content

Marker	Map	Size	n_A	**LG**	**PIC**	**Marker**	Map	Size	n_A	**LG**	**PIC**
ORS543	11.4	268-284	5	1	0.67	ORS691	101.3	375-389	6	10	0.75
ORS716	34.2	317-338	4	1	0.62	ORS621	1.1	252-270	5	11	0.72
ORS837	38.3	447-457	3	1	0.56	ORS457	0.1	242-250	3	11	0.66
ORS610	4.7	157-180	9	1	0.80	ORS1146	49.1	362-398	4	11	0.67
ORS371	45.9	268-276	3	1	0.56	ORS1227	20.3	331-339	5	11	0.71
ORS342	65.3	358-365	3	2	0.24	ORS733	22.4	214-216	4	11	0.56
ORS925	6.5	219-232	7	2	0.77	ORS810	62.5	418-425	2	12	0.45
ORS1065	9.5	290-315	4	2	0.63	ORS1085	71.7	295-298	2	12	0.43
ORS423	1.7	375-393	6	2	0.49	ORS761	48.3	360-368	4	12	0.54
ORS1222	37.7	453-459	3	3	0.63	ORS778	57.7	392-395	3	12	0.31
ORS665	6	304-313	5	3	0.57	ORS502	0.1	111-134	3	12	0.35
ORS949	49.2	372-392	6	3	0.52	ORS630	79	363-370	3	13	0.65
ORS1036	3.1	260-271	2	3	0.50	ORS316	79.9	197-206	4	13	0.53
ORS1114	74.3	257-271	3	3	0.61	ORS1179	60.1	334-339	2	13	0.44
ORS674	100.8	362-374	5	4	0.65	ORS1030	72.1	450-453	2	13	0.26
ORS309	75.5	137-148	2	4	0.47	ORS534	7.1	261-267	5	13	0.73
ORS366	59.6	203-229	5	4	0.68	ORS1248	15.9	388-392	3	14	0.62
ORS505	35.4	250-264	5	5	0.75	ORS1079	14.4	392-414	6	14	0.50
ORS1024	7.7	232-249	7	5	0.70	ORS307	50.2	129-154	3	14	0.52
ORS1120	66.8	311-341	4	5	0.39	ORS832	62.1	353-364	4	14	0.40
ORS852	40.9	217-475	3	5	0.63	ORS694	35.8	180-191	3	14	0.64
ORS483	32.9	285-291	4	6	0.55	ORS687	68.2	178-188	3	15	0.53
ORS381	64.8	229-235	3	6	0.60	ORS857	71.4	227-232	3	15	0.19
ORS1041	17.1	292-300	5	7	0.65	ORS420	0.9	153-159	6	15	0.79
ORS331	24.2	185-198	4	7	0.63	ORS1141	38.6	251-261	5	15	0.75
ORS456	43.9	326-337	3	8	0.51	ORS668	62.1	177-179	2	15	0.17
ORS1161	50.3	239-250	5	8	0.36	ORS656	26.1	217-227	6	16	0.72
ORS894	90	263-273	3	8	0.53	ORS899	0.1	320-341	5	16	0.66
ORS844	75.5	301-326	4	9	0.60	ORS885	95.2	354-357	3	16	0.60
ORS1265	25	205-249	7	9	0.74	ORS750	23	343-359	5	16	0.59
ORS938	10.9	328-340	3	9	0.58	ORS407	71.7	455-480	4	16	0.61
ORS887	38.4	258-266	3	9	0.34	ORS993	44.5	328-344	5	16	0.80
ORS442	110.8	410-424	5	9	0.48	ORS297	29.1	232-243	5	17	0.71
ORS428	18.2	227-235	4	9	0.36	ORS1245	50.8	198-215	5	17	0.62
ORS613	74	218-247	5	10	0.38	ORS561	41	377-449	5	17	0.46
ORS878	29.9	208-221	6	10	0.71	ORS735	80.3	377-391	5	17	0.74
ORS437	59.9	352-362	4	10	0.38						

A total of 295 alleles were amplified, using the 73 primer pairs among the 33 genotypes. The number of alleles per SSR locus varied from 2 to 9, with an average of 4.18. The expected heterozygosity (PIC value) per locus ranged from 0.17 to 0.80, with a mean of 0.56. Genetic distance among the 33 germplasm accessions ranged from 0.02 (KH120R-KH130R) to 0.24 (KH134R-KH141R). The overall mean was 0.591. The evolutionary history was inferred using the UPGMA method (Sneath *et al.*, 1973). The optimal tree with the sum of branch length = 6.90646137 is shown. The tree is drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Phylogenetic analyses were conducted in MEGA4 (Tamura *et al.*, 2007).

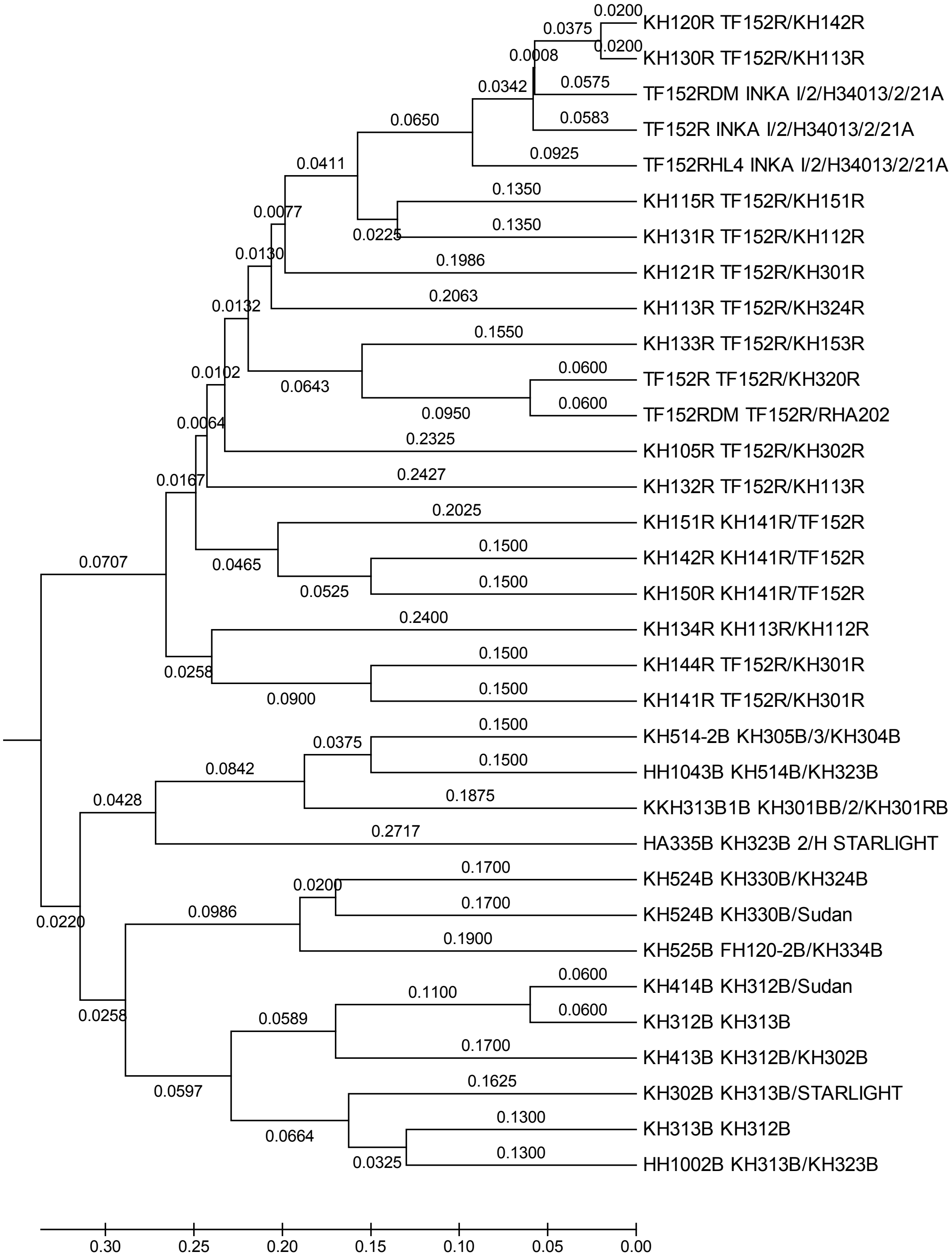

Figure 1. Evolutionary relationships of 33 inbred lines of sunflower.

This method assumes that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. An interesting aspect of this method is that it produces a tree that mimics a species tree, with the branch lengths for two OTUs are the same after their separation. Because of the assumption of a constant rate of evolution, this method produces a rooted tree, though it is possible to remove the root for certain purposes. The algorithm for UPGMA is discussed in detail in Nei and Kumar (2000).

2.5 Discussion

Simple sequence repeats (SSR), also known as microsatellites, are composed of tandem repeated two to six nucleotide DNA core sequences such as (AT)n, (AGC)n, or (GACA)n, and these are spread throughout the genome. The DNA sequences flanking the SSRs are generally conserved within individuals of the same species, allowing the selection of primers that will amplify the intervening SSR. Variation in the number of tandem repeats, results in PCR products of different lengths. SSR markers have the advantage of being highly polymorphic, co-dominant, abundant, and rapid and technically simple to test for, thus they are widely used for DNA fingerprinting and genetic mapping.

The mean number of alleles and the mean PIC values obtained in this study were similar to those reported by Paniego *et al.* (2002), Yu *et al.* (2002), Tang and Knapp (2003), for different sets of sunflower inbred lines. Based on the PIC values of the markers, it is clear that not all 73 SSRs have the same efficiency for routine genotyping and variety identification in this set of sunflower inbred lines.

The clustering method used was the unweighted pair group with arithmetic average clustering (UPGMA; Sneath and Sokal, 1973). The dendrogram constructed using the data derived from all the 73 SSRs grouped the 33 genotypes into two major clusters. The first major cluster consisted of all the R –lines (with a genetic mean of 0.42). The second major cluster consisted of the M (B)-Lines (with a genetic mean of 0.52). This means that the inbred lines used in this study were original lines. The lowest genetic distance values were observed between particular pairs of lines. The average among the isogenic TF152R lines was 0.313. Theoretically, the only difference between each pair of isogenic lines is suppose to be either the gene responsible for the downy mildew resistance, or the quantitative gene effect of high oleic acid that was not covered by the set of SSR used. However, there is a residual heterogeneity between isogenic lines after the backcross procedure and the above genetic distance reflected this. If looking at the downy mildew resistant and susceptible version of the same line, the relative small genetic distance is quite important and could most likely be explained by the number of backcrosses that was probably too limited to reduce the genetic background of the non-recurrent parent. The relative large genetic distance between the isogenic normal and high oleic inbred line is also significant and is most likely due to the screening method employed for the oleic acid content and the number of backcross cycles.

Table 2. Number of alleles and genetic diversity for the two subsets of lines (B and R) of sunflower (*Helianthus annuus*)

	No. alleles			**Genetic diversity**		
Population	Mean	Min	Max	Mean	Min	Max
M lines	3.09	2	7	0.52	0.12	0.78
R lines	2.68	1	5	0.42	0.04	0.72

In terms of gene diversity and allelic richness (i.e., number of alleles per locus), similar results were obtained within each group of lines. Zhang *et al.* (1995) described the distribution of genetic diversity within and between populations that showed that a large proportion of the total diversity was maintained within each group of maintainer and restorer lines, respectively. Overall results of this genetic diversity study showed remarkable correlations with the pedigree information available on this set of inbred lines. Clear traces of the inbred lines used in previous line development could be seen in the dendrogram. Some slight deviations could easily be explained by looking at high resolution ultrathin iso-electric focusing gel protein profiles that identified different allele forms in some inbred lines of sunflower. This could be due to the continuous improvement to the plant material by the breeder. When a new variety is introduced, a reference seed lot is supplied to the Genomics Laboratory and subsequent submissions are compared to the profile of the reference seed. Hence the lab should be able to pick up small differences when they occur. The level of heterogeneity observed in this study was low, suggesting that the cultivated sunflower inbred lines were correctly fixed. Total protein analysis performed on the same lines suggested a level of heterogeneity at the molecular level for some inbred lines. This can be explained by the fact that the selection of sunflower inbred lines is solely based on phenotypic traits.

In this study a relatively large number of SSRs were used to generate diversity results. There was a clear split between the Restorer and the Maintainer lines. There was a similar level of genetic diversity maintained in each genetic pool. South African sunflower breeders may use these results to choose parental lines to maximize variability among lines.

2.6 References

Aguade´, M., 2001. Nucleotide sequence variation at two genes of the phenylpropanoid pathway, the FAH1 and F3H genes in *Arabidopsis thaliana*. Molecular and Biological Evolution 18: 1–9.

Anderson, J.A., Churchill, G.A., Autrique, J.E., Tanksley, S.D. and Sorrells, M.E. 1993. Optimizing parental selection for genetic linkage maps. Genome. 36: 181-186.

Arias, D.M. and Reiseberg, L.H. 1995. Genetic relationship among domesticated and wild sunflower (*Helianthus annuus*, Asteraceae). Economic Botany 49: 239-248.

Aksyonov, I.V. 2005. Use of albumin markers for defining genetic purity of sunflower parent lines and hybrids. Helia 28: 43-48.

Aksyonov, I.V. 2005. Protein markers specificity of sunflower inbred lines. Helia 29: 49-54.

Burke, J.M and Lui, A. 2006. Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics 173: 321–330.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H. and Neale, D.B. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. Procedures of the National Academy of Science USA 101: 15255–15260.

Carrera, A.D., Pizarro, G., Poverene, M., Feingold, S., León, A.J. and Berry, S.T. 2002. Variability among inbred lines and RFLP mapping of sunflower isozymes. Genetics and Molecular Biology 25: 65-72.

Carrera, A., Poverene, M. and Rodriguez, R.H. 1996. Isozyme variability in *Helianthus argophyllus*. Its application in crosses with cultivated sunflower. Helia 19: 19-28.

Cregan, P.B., Jarvik, T., Bush, A.L., Shoemaker, R.C., Lark, K.G., Kahler, A.I., Kaya, N., Van Toai, T.T., Lohnes, D.G., Chung, J. and Specht, J.E. 1999. An integrated genetic linkage map of the soybean genome. Crop Science 39: 1464-1490.

Cronn, R., Brothers, M., Klier, K., Bretting, P.K. and Wendel, J.F. 1997. Allozyme variation in domesticated annual sunflower and its wild relatives. Theoretical and Applied Genetics 95: 532-545.

Garcı´a-Gil, M.L., Mikkonen, M. and Savolainen, O. 2003. Nucleotide diversity at two phytochrome loci along a latitudinal cline in *Pinus sylvestris*. Molecular Ecology 12: 1195–1206.

Garris, A.J., McCouch, S.R. and Kresovich, S. 2003 Population structure and its effects on haplotype diversity and linkage disequilibrium surrounding the xa5 locus of rice (*Oryza sativa* L.). Genetics 165: 759–769.

Gedil, M. A., Wye, C., Berry, S., Segers, B., Peleman, J., Jones, R., Leon, A., Slabaugh, M.B. and Knapp, S.J. 2001. An integrated restriction fragment length polymorphism-amplified fragment length polymorphism linkage map for cultivated sunflower. Genome 44: 213-221.

Gentzbittel, L., Mestries, E., Mouzeyar, S., Mazeyrat, F., Badaoui, S., Vear, F., Tourvieill de Labrouhe, D. and Nicolas, P. 1999. A composite map of expressed sequences and phenotypic traits of the sunflower (*Helianthus annuus* L.) genome. Theoretical and Applied Genetics 99: 218-234.

Hamblin, M.T., Mitchell, S.E., White, G.M., Gallego, J., Kukatla, R. 2004. Comparative population genetics of the panicoid grasses: sequence polymorphism, linkage disequilibrium and selection in a diverse sample of *Sorghum bicolor*. Genetics 167: 471–483.

Hongtrakul, V., Huestis, G.M. and Knapp, S.J. 1997. Amplified fragment length polymorphisms as a tool for DNA fingerprinting sunflower germplasm: genetic diversity among oilseed inbred lines. Theoretical and Applied Genetics 95: 400-407.

Ingvarsson, P.K. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus termula* L., Salicaceae). Genetics 169: 945–953.

Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D. G. Neil, D.L. and Armour, J.A.L. 1994. Complex gene conversion events in germline mutations at human minisatellites. Nature Genetics 6: 136-145.

Kado, T., Yoshimaru, H., Tsumura, Y. and Tachida, H. 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae *sensu lato*). Genetics 164: 1547–1599.

Mitchell, S.E., Kresovich, S., Jester, C.A., Hernandez, C.J. and Szewe-McFadden, A.K. 1997. Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Science 37: 617-624.

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

Nei, M. and Kumar, S. 2000. Molecular Evolution and Phylogenetics. Oxford University Press, New York.

Nordborg, M., Borevitz, J.O., Bergelson,J., Berry, C.C., Chory, J. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. Nature Genetics 30: 190–193.

Paniego, N., Echaide, M., Munoz, M., Fernandez, L., Torales, S., Faccio, P., Fuxan, I., Carrera, M., Zandomeni, R., Suarez, E.Y., and Hopp, H.E. 2002. Microsatellite isolation and characterization in sunflower (*Helianthus annuus* L.). Genome 45: 34-43.

Rachid Al-Chaarani, G., Roustaee, A., Gentzbittel, L., Modrani, L., Barrault, G., Dechamp-Guillaume, G. and Sarrafi, A. 2002. A QTL analysis of sunflower partial resistance to downy mildew (*Plasmopara halstedii*) and black stem (*Phoma macdonaldii*) by the use of recombinant inbred lines (RILs). Theoretical and Applied Genetics 104: 490-496.

Ramos-Onsins, S.E., Stranger, B. E., Mitchell-Olds, T. and Aguade´, M. 2004 Multilocus analysis of variation and specialization in the closely related species *Arabidopsis halleri* and *A. lyrata*. Genetics 166: 373–388.

Reiseberg, L., Baird, S.J.E. and Desrochers, A. M. 1998. Patterns of mating in wild sunflower hybrid zones. Evolution 52: 713-726.

Savolainen, O., Langley, C.H., Lazzaro, B.P. and Fre´ville, H. 2000 Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Molecular Biological Evolution 17: 645–655.

Senior, M.L., Murohy, M.M., Goodman, M.M., Stuber, C.W. 1998. Utility of SSRs for determining genetic similarities and relationships in maize using agarose gel systems. Crop Science 38: 1088-1098.

Sneath, P.H.A and Sokal, R.R. 1973 Numerical Taxonomy. Freeman, San Francisco.

Tamura, K. Dudley, J., Nei, M. and Kumar, S. 2007 MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24:1596-1599. (Publication PDF at http://www.kumarlab.net/publications)

Tang, S., Yu, J.K., Slabaugh, M.B., Shintani, D.K. and Knapp, S.J. 2002. Simple sequence repeat map of the sunflower genome. Theoretical and Applied Genetics 105: 1124-1136.

Tang, S., Kishore, V.K. and Knapp, S.J. 2003. PCR-multiplexes for genome-wide framework of simple sequence repeat marker loci in cultivated sunflower. Theoretical and Applied Genetics 107: 6-19.

Tenaillon, M.I., Sawkins, M.C., Anderson, L.K., Stack, S. M. and Doebley, J. 2002. Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Genetics 162: 1401–1413.

Van Berloo, R. 2007. GGT graphical genotypes. Laboratory of Plant Breeding Wageningen University. The Netherlands. (http://www.dpw.wau.nl/pv/pub/ggt/)

White, S.E., and Doebley, J.F. 1999. The molecular evolution of terminal ear1, a regulatory gene in the genus *Zea*. Genetics 153: 1455–1462.

Wright, S., Lauga, B. and Charlesworth ,D. 2003. Subdivision and haplotype structure in natural populations of *Arabidopsis lyrata*. Molecular Ecology 12: 1247–1263.

Yu, J.K., Mangor, J., Thompson, L., Edwards, K.J., Slabaugh, M.B. and Knapp, S.J. 2002. Allelic diversity of simple sequence repeat markers among elite inbred lines in cultivated sunflower. Genome 45: 652-660.

Yu, J., Tang, S., Slabaugh, M.B., Heesacker, A., Cole, G., Herring, M., Soper, J., Han, F., Chu, W., Webb, D.M., Thompson, L., Edwards, K.J., Berry, S., Leon, A.J., Grondona, M., Olungu, C., Maes, N. and Knapp, S.J. 2003. Towards a saturated molecular genetic linkage map for cultivated sunflower. Crop Science 43: 367-387.

Zhang, Y.X., Gentzbittel, L., Vear, F., and Nicolas, P. 1995. Assessment of inter- and intra-inbred line variability in sunflower (*Helianthus annuus*) by RFLP. Genome 38: 1040-1048.

Zhang, L.S., Le Clerc, V., Li, S. and Zhang, D. 2005. Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. Canadian Journal of Botany 83: 66-72.

Zhu, Y.L., Song, Q.J., Hyten, D.L., Van Tassell, C.P. and Matukumalli, L.K. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163: 1123– 134.

CHAPTER 3: A sub-set of SSR markers for tailed-multiplex analysis and fingerprinting of sunflower inbred lines

3.1 Abstract

An understanding of genetic diversity among parental lines would be useful in hybrid sunflower breeding. Simple sequence repeat (SSR) markers could be used as the molecular markers for an investigation of parental lines of sunflower. Among the different classes of molecular markers, SSRs are the most useful because of their high polymorphism, random distribution, co-dominant Mendelian inheritance and high mutation rate. The objective of this study was to simplify the procedure and to reduce the cost of fluorescent SSR analysis through the identification of (i) a core set of SSRs and (ii) the multiplexing of selected SSR markers, through the tailed primer strategy. Outstanding single-locus SSR markers in the set of sunflower inbreds used for this study were identified. The selected markers produced robust PCR products, amplified a single locus each, were polymorphic among the elite inbred lines and supplied a good, genome-wide framework of completely co-dominant, single-locus DNA markers for molecular breeding. The use of a fluorescent-tailed primer technique resulted in a considerable cost saving. Furthermore, the SSR markers can be multiplexed through optimization, in order to avoid undesirable primer-primer interactions and non-specific amplification.

3.2 Introduction

Sunflower (*Helianthus annuus* Linnaeus) is one of the four major oilseed crops in the world. In the last decade sunflower has been the subject of intense molecular genetics and genomic studies (Hvarleva *et al.*, 2007). The use of SSR markers to assist with breeding through the molecular characterization and identification of plant genotypes has become an important tool for plant breeders. Optimizing the system of molecular markers for sunflower could offer an improvement to the efficiency and affordability of sunflower variety testing. Methods to improve the speed and efficiency of SSR genotyping are integral to the application of molecular markers in plant breeding and research (Hayden *et al*., 2008).

Multiplex PCR (Chamberlain *et al*., 1988) is a variation of the PCR technique used for applications where it is advantageous to amplify two or more loci simultaneously in the same reaction. It is usually used to increase the amount of information generated per assay, and to reduce the use of consumables, time and labour costs (Henegariu *et al*., 1997). This technique usually requires extensive optimization. The widespread use of multiplex PCR for SSR genotyping in crop plants has been limited by several factors. Firstly, PCR multiplexes have been developed for a limited number of SSR markers on a very limited number of crops. But for most crops, no PCR multiplexes have been developed (Liu *et al.,* 2000; Gethi *et al.,* 2002). Secondly, the number of polymorphic SSR marker loci required for molecular breeding applications is often more than the number used in the multiplex PCR reactions. Thirdly, some SSR primers and primer combinations are recalcitrant to being used in a multiplex PCR.

PCR-multiplexing are ideal for genotyping where common sets of SSR marker loci are required for repetitive DNA fingerprinting of new inbred lines and for fast inbred identification. The role of various ingredients in the multiplex PCR, and the protocols for several multiplex PCR techniques have been described by several research groups (Henegariu *et al*., 1997; Zhang *et al*., 2003). The two largest obstacles to successful multiplex PCR are undesirable primer-primer interactions, and non-specific amplification (Elnifro *et al*., 2000). A third obstacle is that the use of a tailed forward primer and a standard length reverse primer in the M13-tailed primer method can promote the amplification of non-specific DNA products. Therefore, the PCR conditions required for amplification using the M13- tailed primer method are often different to those that are optimal for amplification using standard length primers.

The M13-tailed primer method (Oetting *et al.*, 1995) is mostly used for the assay of SSRs, in order to reduce the cost of fluorescent primer labelling, which can be 5-10 times more expensive than the synthesis of an unlabeled primer. This method uses a three primer approach. A PCR is performed using a forward primer with a nucleotide extension at its 5'-end, identical to the sequence of an M13 sequencing primer (5'-CACGACGTTGTAAAACGAC-3'), a standard length reverse primer and a fluorescently labelled M13 primer. During PCR, the SSR product is fluorescently labelled following participation of the M13 primer after the first few cycles of amplification. Thus, instead of synthesizing one specific fluorescently labelled primer for each SSR marker, the labelled M13 primer is the sole source of label. It can be used with any primer that contain the same sequence tail, and generates a labelled amplified DNA fragment.

Fluorescently labelling of SSR markers for genotyping on automated sequencers has many advantages over earlier techniques that used auto-radiographic or silver-stained detection techniques. Firstly, a large increase in throughput is made possible by the multiplexing of many PCR products into a single lane. Secondly, there is a significant increase in the accuracy of allele sizing, achieved by the use of an internal size standard in each lane, combined with automated allele-calling algorithms. Thirdly, it is much quicker than conventional gel systems. Overall, automating the process increases the speed and accuracy of data collection and processing. The high sensitivity of detection also reduces the minimum volume of the PCR reaction, reducing its costs. Its sensitivity also allows for the detection of loci that are difficult to amplify.

Carrano *et al.,* (1989) first reported on the use of fluorescence-based semi-automated analysis of marker panels. This method was adapted and improved upon for SSR analysis by Ziegle *et al.*, 1992. Semi-automated methods of SSR genotyping have gradually replaced manual systems in plant breeding and genetics research. These methods facilitate the efficient application of microsatellite markers for high-throughput mapping (Tang *et al.*, 2002; Zhang *et al.*, 2005), pedigree analysis (Lexer *et al.*, 1999), fingerprinting of accessions (Carrano *et al.,* 1989), and assaying genetic diversity (Macaulay *et al.,* 2001, Zhang *et al.*, 2005). The technology potentially has multiple applications. It can improve the efficiency of managing a germplasm collection, help deliver purity-proven seed stocks to growers, and provide the basis of intellectual property protection (Mitchell *et al.*, 1997). The purpose of this project was to develop and apply multiplex panels of fluorescently labelled microsatellite markers for semi-automated genotyping of *H. annuus* at the whole genome level.

3.2.1 PCR

The use of a sequencer necessitates the use of fluorescent labelled DNA fragments. The most common practice is to label DNA fragments by incorporating the dye into a PCR product using a labelled primer. Fluorescent labelled primers are expensive, especially when used for genotyping projects that involve the use of large numbers of SSR markers. A cost effective alternative is the use of M13 tailed method. The M13 primer sequence (5'-CACGACGTTGTAAAACGAC-3') was added as a standard "tail" to the 5' end of the forward primer during primer synthesis.

Amplification thus needs the presence of three primers: a forward primer with the tail, a reverse non-tailed primer and a fluorescent dye labelled M13 primer. The labelled M13 primer is the only source of label and could be used with any primer that contains the same sequence as the tail to generate a labelled fragment.

Within a single amplification reaction the PCR amplification occurs in two stages. Amplicon 1 is produced using only the tailed forward and the 3' reverse primer, the extension of the forward primer yields a product that contains the "tail sequence". Thus when this template anneals with the reverse primer and extends, a product containing the complement of the tail sequence is produced (amplicon 2). The final step is the production of amplicon 3 by using the labelled M13 primer and amplicon 2 as template. The fluorescent reporter was incorporated into the product during polymerization and a fluorescent signal was emitted. The DNA sequencer will only detect the labelled amplicon 3. (Figure 1.)

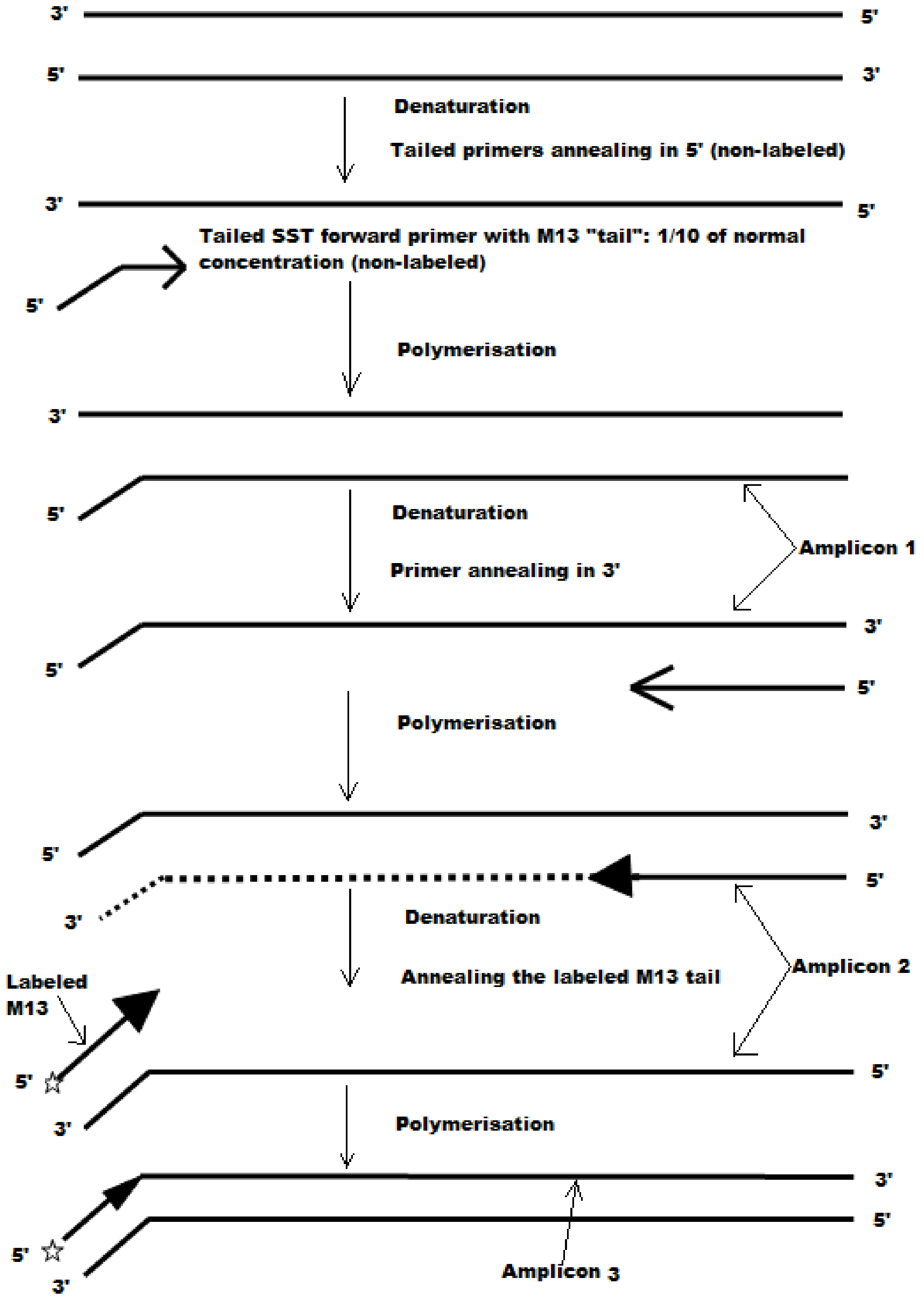

Figure 1. Tailed primer strategy (Zhang *et al.,* 2003)

3.2.2 Core set identification

The efficient identification of genotypes is dependent on the optimum quantity of loci with the maximum number of alleles with clear readability, and according to Antonova *et al.,* 2006, using 1 – 3 molecular markers per chromosome is optimal for the molecular genetic characterisation of cultivated varieties. In 2004 Hlestkina *et al.,* concluded that an increase in the number of microsatellite markers for one variety only leads to a more detailed molecular genetic description of the sample being studies but does not influence the efficiency of the identification.

The polymorphism information content (PIC) for each SSR marker was determined as described by Smith *et al.,* (1997). PIC is a measure of allele diversity at a locus and is equal to:

$$1 - \sum_{i=in}^{n} f_i^2$$

Where fi is the frequency of the i th allele. The PIC, when calculated like this, is synonymous with the term “gene diversity”, as described by Senior *et al.,* (1998). The PIC values provides an estimate of the discriminatory power of a marker looking at the number of alleles at a locus, but also at the relative frequencies of those alleles in the samples being studied. Therefore marker loci with a large number of alleles occurring at equal frequencies will have the highest PIC values.

3.3 Materials and Methods

3.3.1 DNA Extraction

Genomic DNA was extracted from 7 day old seedlings of 33 inbred sunflower lines, grown under controlled conditions. Five individuals per germplasm accession were harvested. Approximately 400mg of young leaf tissue was harvested into a mortar and ground under liquid nitrogen. A 100mg of the frozen ground leaf material were weighed into an eppendorf and the DNA extracted using a Sigma Nucleic Extraction kit, according to the supplier's specifications. The DNA concentration was determined using 0.7% TBE agarose. A working concentration of 10ng µl-1 was standardized for all extracted DNA.

3.3.2 Developing and testing PCR-multiplexes

The criteria used to select SSR markers for PCR-multiplexing were:

a. primer compatibility
b. genotype performance when amplified by multiplex PCR
c. allele length range, map position and heterozygosity.

The SSR markers were sorted by allele-length range and combined to minimize the co-migration of identically labelled non-allelic bands. The goal was to identify at least 4 - 5 SSRs per multiplex, based on tail labelling and minimum injections. Thus 4 - 5 markers were amplified per PCR. The compatibilities of different SSR primer combinations were tested and assessed by screening four public lines.

Twenty six of the 73 markers screened were chosen for developing 2 multiplex sets of five markers each and 4 multiplex sets of four markers each, based on:

a. level of polymorphism detected in screened varieties
b. compatible allele size range
c. similar optimal reaction conditions

d. ease of score.

Firstly, single primer PCR amplifications were performed to check the primer set ease of score using fluorescent dyes, and to compare resultant reproducibility in single and multiplex reactions. Single primer PCR reactions were performed in 12µl of reaction mixture containing 1 x PCR buffer, 2.5mM Mg++, 0.2µl each of dNTPs (Bioline), 1 unit of Taq polymerase (Bioline) and 5 - 10ng of genomic DNA. Primers were labelled with a fluorescent dye; using the tailed primer strategy (Zhang *et al.,* 2005), one tailed M13 (5'-CACGACGTTGTAAAACGAC-3'), was added to 5'-end of one of the SSR primers (forward primer) during primer synthesis. Three primers are required for the amplification of each SSR locus:

a. one tailed forward primer (0.025µmol)
b. one normal reverse primer (0.25µmol)
c. one labelled tail (0.25µmol).

On the ABI 3130xl Sequencer, a dye set consisting of 5 different dyes were chosen:

a. FAM (Blue)
b. VIC (Green)
c. NED (Yellow)
d. PET (Red)
e. LIZ (Orange – this colour was used for the internal LIZ-size standard).

A "Touchdown" PCR was used to reduce spurious amplification. The initial denaturation step was performed at 94ºC for 2min, followed by 1 cycle at 94ºC for 30s, 63ºC for 30s and 72ºC for 45s. The annealing temperature was decreased by 1ºC per cycle in subsequent cycles until reaching a temperature of 57ºC. Products were subsequently amplified for 32 cycles at 94ºC for 30s, 57ºC for 30s, and 72ºC for 45s, with a final extension for 20 min. Amplifications were performed using a GeneAmpPCR System 9700 (Applied Biosystems) thermal cycler.

To optimise multiplex reactions, the first primers were added in equal amounts in the multiplex PCR reaction. Concentrations were then optimised according to the level of amplification observed for each marker at a particular concentration, aimed at obtaining a similar level of amplification in each multiplex set (Henegariu *et al.,* 1997).

Amplified loci were detected by laser scanning during electrophoresis, using an ABI 3130xl Sequencer (Applied Biosystems). Samples containing 1µl of the PCR products were mixed with 8.5µl loading buffer (formamide) and 0.5µl of the Liz-250 internal standard (ABI). Samples were denatured at 95ºC for 5min, cooled to 4ºC, then loaded on an auto-sampler for auto-injection and capillary electrophoresis. Band sizes were generated automatically in comparison with a standard sizing ladder, included in every sample prior to electrophoresis, using Genescan® and Genotyper® computer software. Band scoring was then checked manually. Banding-profile reproducibility was assessed by repeating experiments in independent single and multiplex PCRs and electrophoreses, using bulked DNA samples.

Multiplex PCR was performed in 15µl of a reaction mixture containing 0.8x PCR buffer, 2.5mM Mg++, 0.2µl each of dNTPs(Bioline), 1 unit of Taq polymerase (Bioline) and 10ng of genomic DNA. Primers were labelled with a fluorescent dye; using a Tailed Primer Strategy (Zhang *et al.,* 2005), one tailed M13 (5'-CACGACGTTGTAAAACGAC-3'), was added to 5'-end of one of the SSR primers (forward primer) during primer synthesis. For a four primer multiplex, nine primers were required for the amplification of each SSR locus: four tailed forward primers (0.025µmol to 0.062µmol of each), four normal reverse primers (0.25µmol to 0.625µmol of each) and one labelled tail (0.25µmol to 0.625µmol).

A "Touchdown" PCR was used to reduce spurious amplification. The initial denaturation step was performed at 94ºC for 2min, followed by 1 cycle at 94ºC for 30s, 63ºC for 30s and 72ºC for 45s. The annealing temperature was decreased by 1ºC per cycle in subsequent cycles until reaching a temperature of 57ºC. Products were subsequently amplified for 32 cycles at 94ºC for 30s, 57ºC for 30s, and 72ºC for 45s with a final extension for 20 min. Amplifications were performed using a GeneAmpPCR System 9700 (Applied Biosystems) thermal cycler.

Amplified loci were detected by laser scanning during electrophoresis, using an ABI 3130xl Sequencer (Applied Biosystems). Samples containing 1µl of the PCR products were mixed with 8.5µl loading buffer (formamide) and a 0.5µl Liz-250 internal standard (ABI). Samples were denatured at 95ºC for 5min, cooled to 4ºC, then loaded on the auto-sampler for auto-injection and capillary electrophoresis. Band sizes were generated automatically in comparison with a standard sizing ladder included in every sample prior to electrophoresis, using Genescan® and Genotyper® computer software. Band scoring was then checked manually. Banding-profile reproducibility was assessed by repeating experiments in independent single and multiplex PCRs and electrophoreses using bulked DNA samples.

3.3.3 Pooling PCR Reactions

The optimal pooling of PCR reactions is determined by the dye set chosen. For the dye set used in this study, up to four separate PCR reactions could be pooled into one ABI sample, if each PCR multiplex used a different M13 dye. The pooling of samples greatly reduces costs and increase throughput. The four sample dyes all fluoresce at different wavelengths and different intensities. To overcome the intensity differences, the different reactions have to be pooled at different pooling ratios. The pooling ratio followed consisted of 3.0µl FAM : 3.0µl VIC : 4.0 µl NED : 6.0µl PET, placed into 14µl of water.

The samples were suspended in formamide to denature the DNA. 0.15µl of the LIZ-250 size standard was added to 9.85 µl Hi-Di formamide and 3.0 µl of the pooled samples was added for a final volume of 13 µl per sample. The samples were denatured at 95ºC for 5 minutes and immediately cooled to 4ºC and loaded onto the auto-sampler for auto-injection and capillary electrophoresis. Band sizes were generated automatically in comparison with a standard sizing ladder included in every sample prior to electrophoresis, using Genescan® and Genotyper® computer software.

The primer - primer interactions are usually difficult to manage during multiplex optimization. The levels of primer – primer interaction were therefore evaluated during multiplexing optimization using a software package from “FastPCR”.

3.4 Results

In Table 1, allele numbers and other summary statistics are reported for the SSR markers that were selected as the core set.

Table 1. A summary of the chosen core set of SSRs.

Marker	Map	Size	nA	LG	PIC	Sets	Primer Con. (µmol L-1)
ORS543	11.4	268-284	5	1	0.67	1	0.375
ORS610	4.7	157-180	9	1	0.80	1	0.25
ORS366	59.6	203-229	5	4	0.68	1	0.25
ORS1141	38.6	251-261	5	15	0.75	1	0.25
ORS925	6.5	219-232	7	2	0.77	2	0.25
ORS505	35.4	250-264	5	5	0.75	2	0.25
ORS691	101.3	375-389	6	10	0.75	2	0.375
ORS993	44.5	328-344	5	16	0.80	2	0.375
ORS1265	25	205-249	7	9	0.74	3	0.25
ORS534	7.1	261-267	5	13	0.73	3	0.375
ORS1248	15.9	388-392	3	14	0.62	3	0.625
ORS694	35.8	180-191	3	14	0.64	3	0.25
ORS420	0.9	153-159	6	15	0.79	3	0.25
ORS1222	37.7	453-459	3	3	0.63	4	0.375
ORS381	64.8	229-235	3	6	0.60	4	0.25
ORS878	29.9	208-221	6	10	0.71	4	0.25
ORS735	80.3	377-391	5	17	0.74	4	0.372
ORS1065	9.5	290-315	4	2	0.63	5	0.625
ORS1024	7.7	232-249	7	5	0.70	5	0.25
ORS621	1.1	252-270	5	11	0.72	5	0.25
ORS1227	20.3	331-339	5	11	0.71	5	0.625
ORS656	26.1	217-227	6	16	0.72	5	0.25
ORS1041	17.1	292-300	5	7	0.65	6	0.375
ORS894	90	263-273	3	8	0.53	6	0.25
ORS630	79	363-370	3	13	0.65	6	0.375
ORS297	29.1	232-243	5	17	0.71	6	0.25

3.5 Discussion

Multiplex optimization required the combination of primers in various mixes, because of the amplification of many loci at the same time. For the first amplification of the multiplex samples, equimolar amounts of all the primers were used. The multiplex PCR of four and five loci often lead to uneven amplification efficiency of the PCR products. Longer loci with sizes over 350bp for example ORS691: 375bp, ORS1248: 388bp, ORS1222: 453bp, ORS735: 377bp and ORS630: 363bp showed lower amplification efficiency yields.

All multiplex sets were tested for dimmer formation between and among all primers using specific software, to exclude primer - primer interaction as reason for low product formation. The only competition in the multiplex reactions was for the limited amount of enzyme and nucleotides.

An increase in primer concentration of the primers with long loci products increased the yield and visibility of these loci substantially. The increase of primer concentrations of only the longest loci primers of the five primers used in multiplexing resulted in some suppression on the second longest loci's efficiency. The increase in primer concentration of this primer also led to optimal amplification of all four primers in the four primer multiplexes.

In the five primer multiplexes, it was necessary to increase the primer concentration of both larger loci primers to an even higher concentration than in the four primer multiplex set. The concentrations of the primers in the sets are listed in Table 1.

A further optimization was performed on some of the components of the PCR mix. The concentration of the buffer was adjusted from a 1x to 0.8x concentration. This adjustment had the greatest effect on longer amplification products because lower salt concentrations favour larger products, and higher salt concentrations favour shorter amplification products. The $MgCl_2$

concentration was kept unchanged. Changes to the template and the Taq polymerase concentration made no significant difference to the efficiency of the multiplex reaction.

A twenty six core set was developed that was successful in discriminating between all the inbred lines used in this study. It could determine the genetic relationships between varieties, and therefore it could be used for pre-screening and grouping of candidate and existing inbred lines used for producing hybrids. PCR multiplexes for genome-wide or nearly genome-wide collections of SSR marker loci have only been developed for two other plant species thus far, (*Arabidopsis thalianab* Lineaus) (Ponce *et al*., 1999) and maize (Gethi *et al*., 2002)

The primer pairs selected for the multiplex reactions were based on the PIC of the primers, the composition of the primers and the length of the PCR products. Primer - primer interactions were tested using software available from “FastPCR” in order to determine the conditions that minimized interaction levels.

It was essential to test the amplification products of the chosen primers for both single and multiplex reactions because this indicated which primer pair yielded the unspecific products, or failed to produce specific products in the multiplex reaction. Optimization of the different primer concentrations was made easier with the knowledge of each product. The adjustment of the buffer concentration further helped to achieve the optimal amplification of each multiplex reaction.

The ultimate requirement for an optimal multiplex PCR is the amplification of all products without any unspecific by-products, with the use of a universal PCR program that gives optimal results on all multiplex reactions.

The proposed sunflower PCR-multiplexes amplify twice as many SSR marker loci per PCR than the assortment of PCR multiplexes described thus far for maize, cotton and soybean (Liu *et al*., 2000, Narvel *et al.,* 2000, Gethi *et al.*, 2002). The uniqueness of this study lies in the multiplexing using the tailed strategy, whereas all multiplexes in the literature to date have been based on the use of labelled forward primes. The cost and time saving of this new technique are significant. These are summarised in Table 2.

Table 2. Summary of Simplex PCR; Multiplex SSR (Labelled Forward primer) and Tailed Multiplex SSR (per 96 PCRs)

	Simplex PCR	Multiplex PCR	Tailed Multiplex PCR
Time	4 hours per primer pair	4 hours per 6 primer pairs	4 hours per 6 primer pairs
Cost	R 3 033.60	R 2 186.56	R1 886.08

Tailed multiplexing increase genotyping throughput, reduce PCR costs by an estimated 50 to 70% compared to multiple simplex PCRs(Tang *et al.*, 2003), A further cost saving derived from this approach lay in the use of a semi-automated analysis for the final analysis of the PCR products. Amplified loci were detected by laser scanning during electrophoresis, using an ABI 3130xl Sequencer (Applied Biosystems). The four sample dyes used in this system all fluoresce at different wavelengths and different intensities. This feature allows a maximum of twenty loci to be scored from a single analysis of the multiplex sets proposed in Table 1.

3.6 References

Antonova, T.S., Guchetl, S.Z., Tchelustnikova, T.A. and Ramasanova, S.A. 2006. Development of marker system for identification and certification of sunflower lines and hybrids on the basis of SSR-analysis. Helia 29 (45): 63-72.

Carrano, A.V., Lamerdin, J., Ashworth, L.K. Watkins, B., Branscomb, E., Slezak, T., Raff,M., De Jong, P.J., Keith, D., McBride, L., Meister, S. and Kronick, M. 1989. A high resolution, fluorescence based, semi-automated method for DNA fingerprinting. Genomics 4: 129–136.

Chamberlain, J.S., Gibss, R.A., Ranier, J.E., Nguyen, P.N., Caskey, C.T. 1988. Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. Nucleic Acid Research16: 11141-11156.

Coburn, J. R., Temnykh, S. V., Paul, E. M. and McCouch, S. R. 2002. Design and application of microsatellite marker panels for semi-automated genotyping of rice (*Oryza sativa* L.). Crop Science 42: 2092–2099.

Elnifro, E.M., Ashshi, A.M., Cooper, R.J. and Klapper, P.E. 2000. Multiplex PCR: optimisation and application in diagnostic virology. Clinical Microbiology Reviews 13: 559-570.

Gethi, J.G., Labate, J.A., Lamkey, K.R., Smith, M.E. and Kresovich, S. 2002. SSR variation in important U.S. maize inbred lines. Crop Science 42: 951-957.

Hvarleva, T., Bakalova, A., Chepinski, I., Hristova-Cherbadji, M., Hristov, M. and Atanasov, A. 2007. Characterization of Bulgarian sunflower cultivars and inbred lines with microsatellite markers. Biotechnology and Biotechnology Equipment 24: 408-412.

Hayden, M.J., Nguyen, T.M., Waterman, A. and Chalmers, K.J. 2008. Multiplex-Ready PCR: A new method for multiplexed SSR and SNP genotyping. BMC Genomics 9: 80-92.

Henegariu, O., Heerema, N.A., Dlouhy, S.R., Vance, G.H. and Vogt, P.H. 1997. Multiplex PCR: Critical parameters and step-by-step protocol. Biotechniques 23: 504-511.

Hlestkina, E.K., Salina, E.A. and Shumnii, V.K. 2004. Genotyping of soft wheat native cultivars with use of microsatellite (SSR) markers. Agricultural Biology 5: 44-51.

Lexer, C., Heinze, B., Steinkellner,H., Kampfer, S., Ziegenhagen, B. and Glössl , J. 1999. Microsatellite analysis of maternal half-sib families of *Quercus robur*, pedunculate oak: detection of seed contaminations and inference of the seed parents from the off spring. Theoretical and Applied Genetics 99: 185–191.

Liu, S. Cantrell, R.G., McCarty, J.C. and Stewart, J.M. 2000. Simple Sequence Repeat-based assessment of genetic diversity in cotton race stock accessions. Crop Science 40: 1459-1469.

Macaulay, M., Ramsay, L., Powell, W. and Waugh, R. 2001. A representative, highly informative genotyping set of barley SSRs. Theoretical and Applied Genetics 102: 801–809.

Mitchell, S.E., Kresovich, S., Jester, C.A., Hernandez, C.J. and Szewc-McFadden, A.K. 1997. Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Science 37: 617–624.

Narvel, J.M, Chu, W.C., Fehr, W.R. Cregan, P.B. and Shoemaker, R.C. 2000. Development of multiplex sets of simple sequence repeat DNA markers covering the soybean genome. Molecular Breeding 6: 175-183.

Oetting, W.S., Lee, H.K., Flanders, D.J., Wiesner, G.L., Sellers, T.A. and King, R.A. 1995. Linkage analysis with multiplexed short tandem repeat polymorphisms using infrared fluorescence and M13 tailed primers. Genomics 30: 450-458.

Ponce, M.R., Robles, P. and Micol, J.L. 1999. High-throughput genetic mapping in *Arabidopsis thaliana*. Molecular and General Genetics 154: 408-415.

Senior, M.L., Murphy, J.P., Goodman, M.M. and Stuber, C.W. 1998. Utility of SSRs for determining genetic similarities and relationships in maize using an agarose gel system. Crop Science 38: 1088-1098.

Smith, J.S.C. Chi, E.C.L., Shu, H., Smith, O.S., Wall, S.J., Senior, M.L., Mitchell, S.E., Kresovitch, S. and Ziegle, J. 1997. An evaluation of the utility of SSR loci as molecular markers in maize (*Zea mays* L.): comparison with data from RFLPs and pedigree. Theoretical and Applied Genetics 95: 163-173.

Tang, S., Yu, J.K., Slabaugh, M.B., Shintani, D.K. and Knapp, S.J. 2002. Simple sequence repeat map of the sunflower genome. Theoretical and Applied Genetics 105: 1124-1136.

Zhang, L.S., Bacquet, V., Li, S.H. and Zhang, D. 2003. Optimization of multiplex PCR and multiplex gel electrophoresis in sunflower SSR analysis using infrared fluorescence and tailed primers. Acta Botanica Sinica. 45(11): 1312-1318.

Zhang, L.S., Le Clerc, V., Li, S. and Zhang, D. 2005. Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. Canadian Journal of Botany 83: 66-72.

Ziegle, J.S., Su, Y., Corcoran, K.P., Nie,L., Mayrand, P.E., Hoff, L.B., McBide, L.J., Kronick, M.N. and Diehl, S.R. 1992. Application of . automated DNA sizing technology for genotyping microsatellite loci. Genomics 14: 1026–1031.

CHAPTER 4: Development of techniques to remove visual interference of total protein images of sunflower samples, using first stage iso-electric focusing gels

4.1 Abstract

Genetic analysis of hybrid sunflower (*Helianthus annuus* L.) varieties is routinely performed using protein analysis. This is needed for quality control during hybrid sunflower seed production. First stage iso-electric focusing of total protein extracts are often used to analyze sunflower varieties for the purposes of determining their genetic purity, and to conduct genetic variety verification on large numbers of genetically diverse sunflower populations. Severe visual interference often occurs in gels of seed protein extracts of sunflower. This interference often leads to the masking of the inbred markers used during genetic protein purity analyses. Typically, interferences are visible as a distortion in the gel matrix at the anodal end of the gel, causing important proteins to denature in the presence of heightened field strength and the absence of a uniform matrix. The aim of this study was to identify a method to minimize this visual interference.

4.2 Introduction

During sunflower breeding and selection processes, it is essential that genetic purity is controlled. Genetic purity is important for seed companies that guarantee high yielding hybrids as having stable genetics with defined characteristics, such as resistance to certain diseases. Traditionally genetic purity analysis is performed through the use of phenotypic evaluation (Aksyonov, 2005). This typically consists of physical inspections of sunflower plants at various sages of development, the flowering stage being the most important stage to assess purity. Unfortunately, this method has limitations. According to Aksynov (2005), “the morphological parameters are neither sufficiently conspicuous nor sufficiently stable.” Morphological properties are also affected by the environment (Sammour, 1991). According to Zhang *et al*. (2005), “sunflower is a plant that is very sensitive to interactions among genotype, location, and year; the phenotype of the same plant may vary greatly on the same plant material, and may vary according to location and the growing year”. Furthermore, morphologically identical accessions can only be distinguished at a genetic level. Protein electrophoresis is an analytical tool that provides an indirect method for genome probing by exposing structural variations in enzymes and other total proteins (Cooke, 1984).

Electrophoretic markers were believed to be independent of cultivar morphology and physiology (Sammour, 1991). The advantages of using electrophoretic markers for variety and species identification are:

a. they are rapid;
b. they are relatively cheap;
c. they eliminate the need to grow plants to maturity;
d. they are largely unaffected by the environment.

There are some disadvantages, however, in that they are influenced by tissue specificity and developmental stage. This disadvantage can be overcome by using seed storage proteins.

There are typically two classes of plant storage proteins: seed storage proteins (SSPs) and vegetative storage proteins (VSPs). (Fujiwara *et al.,* 2002). SSPs accumulate to high levels in seeds during the late stages of seed development. They are degraded during seed germination and the released amino acids are utilized as a key nutritional resource for the developing seedlings. The SSPs determine the total protein content of the seed and the quality of the seed for end users (Shewry *et al.,* 1995). SSPs account for about 50% of the total protein in mature cereal grains (Shewry *et al.,* 2005). SSP genes are classic targets for plant molecular biology. The high levels of genetic expression of SSP genes in seed allowed for the detection of SSP gene transcripts and cDNA cloning, which took place during the late 1970's to early 1980's (Fujiwara *et al.,* 2002).

Detailed studies of SSPs dates from the turn of the century, when Osborne (1924) classified them into groups on the basis of their extraction and solubility in water (albumins), dilute saline solutions (globulins), alcohol/water mixtures (prolamins), and dilute acids or alkalis (glutelins). The major seed storage proteins include the albumins, globulins and prolamins, according to the "Osborne fractionation". The most recent classification of seed proteins creates three groups: storage proteins, structural and metabolic proteins.

In contrast, Mandal *et al.,* (2000) placed seed proteins into only two basic categories: housekeeping and storage proteins. The housekeeping proteins are responsible for maintaining normal cell metabolism. These proteins are divided into storage, structural and biologically active proteins and the most biologically active proteins are included in this group, i.e., lectins, enzymes and enzyme inhibitors. The SSPs are non-enzymatic and provide a balance of amino acids required during germination and the establishment of a new plant.

A quick overview of the different types of SPPs is necessary for the understanding of the visual interference encountered during electrophoresis and therefore, the proposed solutions. Storage globulins are contained in the embryo and outer aleurone layer of the endosperm. In maize these have been studied in detail by Wallace and Kriz (1991). In sunflower 11S globulin (helianthinin) is a salt soluble protein that is one of the major storage proteins (Anisimova *et al.,* 2004). Prolamin storage proteins are the major endosperm storage proteins of all cereal grains. All individual prolamin polypeptides are alcohol-soluble in the reduced state and vary greatly in molecular weight, from about 10 000 to almost 100 000. Prolamin has an evolutionary and structural relationship to the 2S albumin storage protein (water-soluble) of sunflower.

The execution of total protein genetic purity analysis is usually based on the extraction of a crude protein, followed by a precision separation of the component proteins using a very high resolution ultrathin layer iso-electric focusing (UTLIEF) gel. Visual interference from sunflower seed extractions is primarily due to fats and oils contained at high levels in sunflower seeds. These are co-extracted with the protein of choice, helianthinins or albumins. These are used as molecular markers to distinguish between cultivars, to check species identification, to assist biosystematic analysis and to study phylogenetic relationships of the species (Sammour, 1991).

A protein inbred marker is typically a protein band (one or more proteins) that is expressed in the hybrid and inherited from the inbred male of the hybrid, in the hybrid the protein band is mono-morphic. However, the protein band(s) are polymorphic and absent in the inbred female of the hybrid. Thus the presence of a self-pollinated female will be clearly visible in the hybrid protein electrophoregram because of the absence (polymorphism) of the marker. The analysis of markers allows for the reliable identification of homozygotes (lines) and heterozygotes (hybrids) in sunflower. (Aksynov, 2005) (Figure 1)

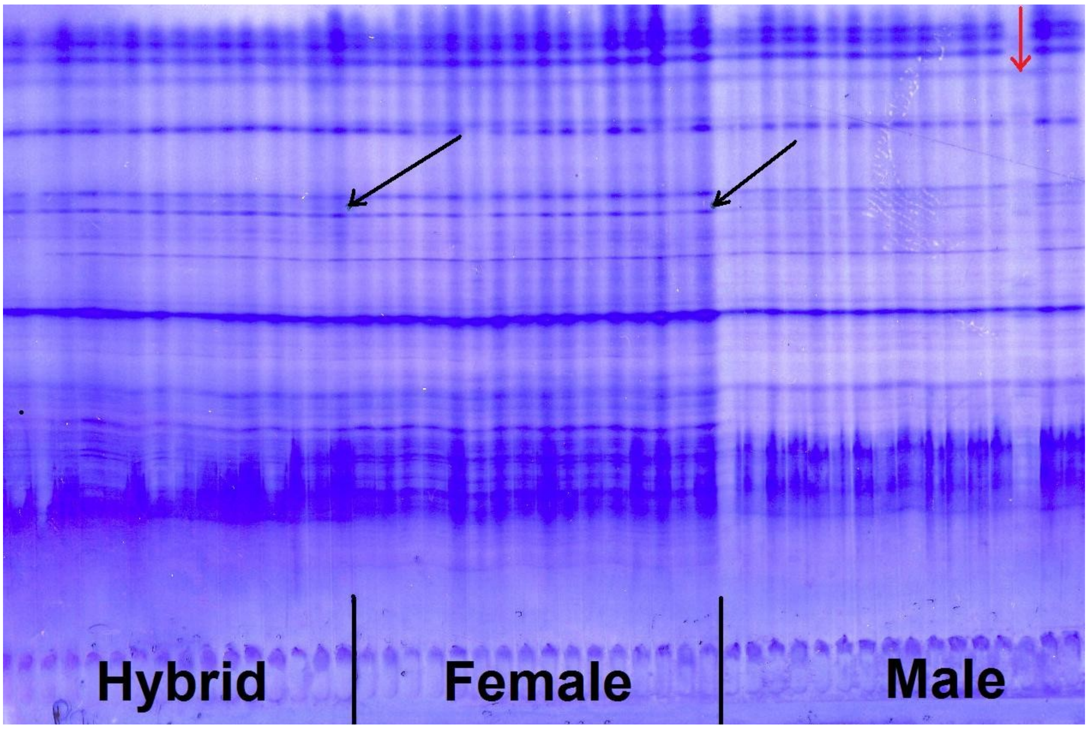

Figure 1. A typical image of a total protein gel. The black arrows indicate the inbred marker. The red arrow indicates a possible off-type in the male inbred seed. Note interference at bottom of gel.

4.3 Materials and Methods

Total protein extractions from seed were performed by distributing the seed to be extracted into 24 well tissue culturing plates, dispensing of the buffer of interest and crushing of the seed in the buffer by the use of an appropriate crushing and mixing apparatus. Extracts were left to imbibe for a minimum of an hour at room temperature. 0.1M Tris-citrate pH7.0 (TC7) and 10% ethanol were used as extraction buffers. The extraction volume was 1ml per kernel.

A Protocols to Reduce the Impact of High Oil Content

In order to determine whether the high oil content of sunflower seeds is the cause of the interference the following alternative extraction protocols were investigated for protein extraction:

a. Normal protein extraction from sunflower seed with TC7
b. Normal protein extraction with TC7 but from de-hulled sunflower seed

c. Normal protein extraction with TC7, using sunflower seed that was squashed onto highly oil-absorbent paper discs prior to extraction into TC7
d. Normal protein extraction with TC7, but using de-hulled sunflower seed that was squashed onto a highly oil-absorbent paper discs prior to extraction into TC7
e. Normal protein extraction with TC7, rapidly freeze extracted protein at -84ºC
f. Normal protein extraction with TC7 crushing de-hulled seed and rapidly freezing the extracted protein at -84ºC
g. Protein extraction into 750µl TC7 + 250µl glacial acetic acid + acetone from stock (stock solution: 3ml glacial acetic acid in 100ml acetone)

B Protocols to Reduce the Impact of Very Large Proteins

To determine whether very large proteins are the cause of the gel interference, the following protocols were tested:

a. Normal protein extraction using TC7
b. Normal protein extraction using TC7, followed by filtering of the extract to remove protein fragments of 1200kDa and bigger.

C Protocols to Reduce the Impact of Phenolic Compounds

The following protocols were tested to determine whether the presence of phenolic compounds could be the cause of the interference:

a. Normal protein extraction using TC7
b. Normal protein extraction using 10% ethanol
c. Protein extraction using 0.1M Tris-citrate pH7 diluted from a 1M Tris-citrate stock, using 10% ethanol as the diluent
d. Protein extraction using 0.1M Tris-citrate pH7 diluted from a 1M Tris-citrate stock, using 30% ethanol as the diluent.

The above extractions were applied to ultra thin iso-electric focusing gels with a wide pI range using large application strips. Pre-focusing was performed on a 12 X 30 PAG Type 1 and Type 2. The gels were supplied by Proteios

International BV. Electrophoresis was performed on a flat bed focuser (Multiphor II electrophoresis system) at a pre-cooled temperature of 10ºC. The anodal buffer consisted of 25.5mM L^{-1} aspartic acid and 24.5mM L^{-1} glutamic acid in distilled water. The cathodal buffer used was 25.2mM L^{-1} arginine, 24.6 mM L^{-1} lysine and 12% ethylenediamine in distilled water. The gels were run using a single cathode and single anode and single direction electrophoresis. The PAG was pre-focused at 200 V, 30W and 12mA for 100 volt hours, using a volt hour integrated electrophoresis power supply (EPS3501 – XL) (Proteios, 2001).

12 µl of each protein extract was loaded individually onto an application strip resting on the gels. Electrophoresis was performed at the following settings: gel entry run at 200V, 30W and 12mA for 100 volt hours and gel focusing at 200V, 30W and 12mA for 1500 volt hours (Proteios, 2001).

After completion of protein focusing, the gels were fixed using 20% tri-chloroacetic acid for 15min without shaking and a further 15min with shaking. The gels were then stained using a standard Coomassie blue stain and a silver stain:

a. Fixing of the proteins in 20% TCA (trichloroacetic acid) solution,
b. Reducing the gel by washing the gels for 3 x 5min in 250ml MAD working solution (The MAD stock solution consisted of 1.5L methanol + 0.75L acetic acid. The working solution was made up with 200ml MAD stock solution, 10 mg dithiothreitol and 800 ml dH_20.). Incubate the gel in 0.1% potassium dichromate solution (prepared immediately before use) for 5min in the dark;
c. Silver stain by incubating the gel for 20min in a 0.2% silver nitrate solution (prepared immediately before use).
d. Develop the gels in 150ml of a sodium carbonate working solution (Stock Solution: 150g sodium carbonate in 1L dH_20; Working Solution: 100ml of stock solution, 400ml dH_20 and 1ml formaldehyde (37%)) for approximately 3min. The solution was changed as soon as it changed colour.

e. Stopping the development by the addition of 1% acetic acid

f. Wash and dry gels.

4.4 Results

A. Protocols to Reduce the Impact of High Oil Content

In the first study, to determine whether high oil content could be the cause of the interference, the following results were obtained:

Table 1. Summary of Results for Different Protein Extraction Protocols in Test A

Test	Description	Buffer	Interference Level
A	Normal	TC7	+++++
B	De-hulled	TC7	++++
C	Oil Pressed	TC7	++++
D	De-hulled + Oil pressed	TC7	++++
E	Normal + Freeze	TC7	+++++
F	De-hulled + Freeze	TC7	++++
G	Normal	TC7 + acetic acid + acetone	+++

B. Protocols to Reduce the Impact of Very Large Proteins

In the second study, to determine whether very large proteins were the cause of the interference, the following results were obtained:

Table 2. Summary of Results for Test B

Test	Description	Buffer	Interference Level
A	Normal	TC7	+++++
B	Normal + Filtered	TC7	+++++

C. Protocols to Reduce the Impact of Phenolic Compounds

In the third study, to determine whether phenolic compounds were the cause of the gel interference, the following results were obtained:

Table 3. Summary of Results for Test C

Test	Description	Buffer	Interference Level
A	Normal	TC7	+++++
B	Normal	10% EtOH	+++++
C	Normal	TC7 +10% EtOH diluent	+++
D	Normal	TC7 +30% EtOH diluent	+++

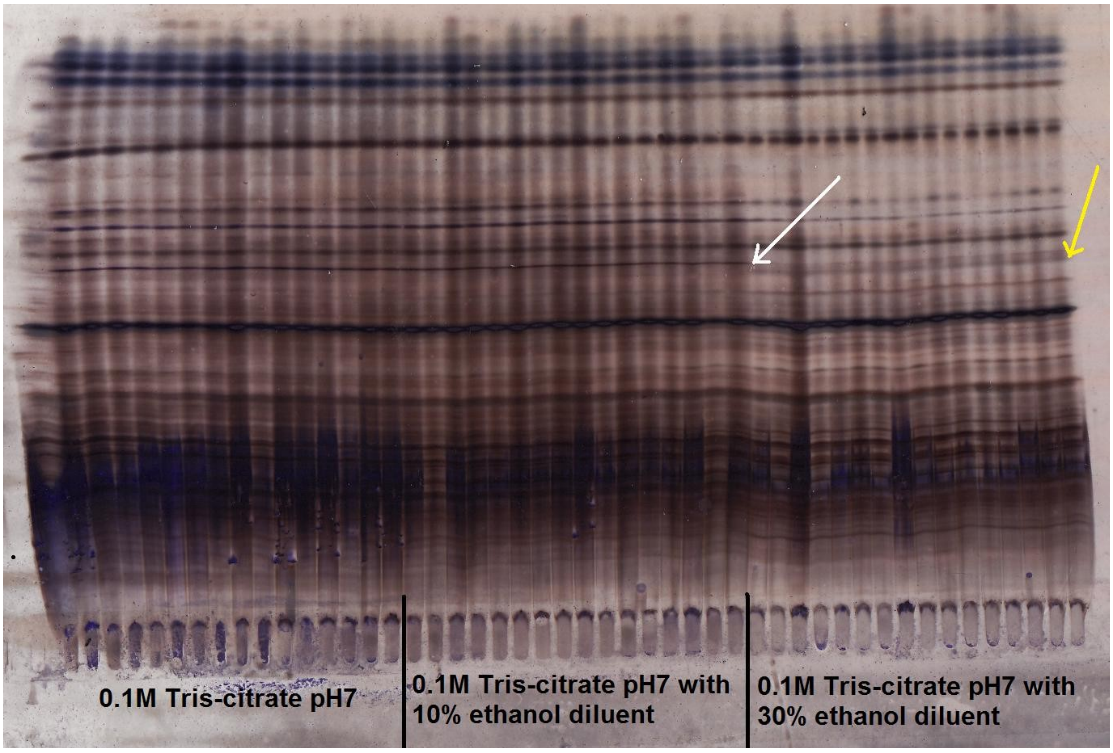
0.1M Tris-citrate pH7
0.1M Tris-citrate pH7 with 10% ethanol diluent
0.1M Tris-citrate pH7 with 30% ethanol diluent

Figure 2. Gel image showing the effect of the %diluent on the protein profile, the white arrow indicate one of the protein bands almost disappearing and the yellow arrow indicate the visibility of a "new" band.

4.5 Discussion

Test A

From the protocols evaluated in Tests A, B and C, none of the protocols alone gave the ultimate solution to visual interference. In Protocol A, the seed was de-hulled to remove a possible source of the interference, the sunflower seed hull consist of lipids, proteins and carbohydrates; lipids represent 5.17% of the total hull weights, 2.96% of which are waxes that are composed of long chain fatty acids (C14–C28, mainly C20) and fatty alcohols (C12–C30, mainly C22, C24, C26). However, the relatively small quantities of lipids and fats contained in the hulls were not the biggest source of the interference. The pressing of seeds onto filter paper was intended to physically remove as much oil as possible from seeds. However, the low effectiveness of these two techniques, even when they were combined, led to the conclusion that physical treatments of the seed were not going to provide solutions to the problem. The quick freezing of the extract was intended to solidify the fats to enable their physical removal, but again, this proved unsuccessful

The chemical removal of the interfering oils and fats were attempted by the addition of a mixture of acetic acid and acetone. Acid hydrolysis has been shown to be a significant contributor to oil degradation at low pH. Therefore acetic acid was added to the extraction buffer. Acetone dissolves oil, and is used in the three phase partitioning of proteins during protein purification. The combination of acetone and acetic acid was a new approach, and therefore empirical trials were needed to identify the correct concentrations of these two solvents, and to determine the correct ratio of acetic acid to acetone (results not shown).

Test B

Protocol B was followed to determine if the problem was caused by large proteins that could not enter the gel matrix and would therefore precipitate out of the gel and tear the matrix during electrophoresis. However, filtering of the extract using small pore size filters made no difference to the gel image.

Test C

A further protocol was tested to determine the impact of phenolic compounds. The test was executed by extracting two different types of proteins, using different solvents. The effect of the interference was less in the ethanol extract that in the TC7 extract. This indicated that phenolics could not be the cause of the interference because phenolics do not dissolve well in ethanol. A combination of extraction protocols was tested to determine whether the two types of proteins could be extracted simultaneously and the occurrence of interference be reduced at the same time. The impact of different levels of ethanol used as a diluent can be seen in Figure 3.

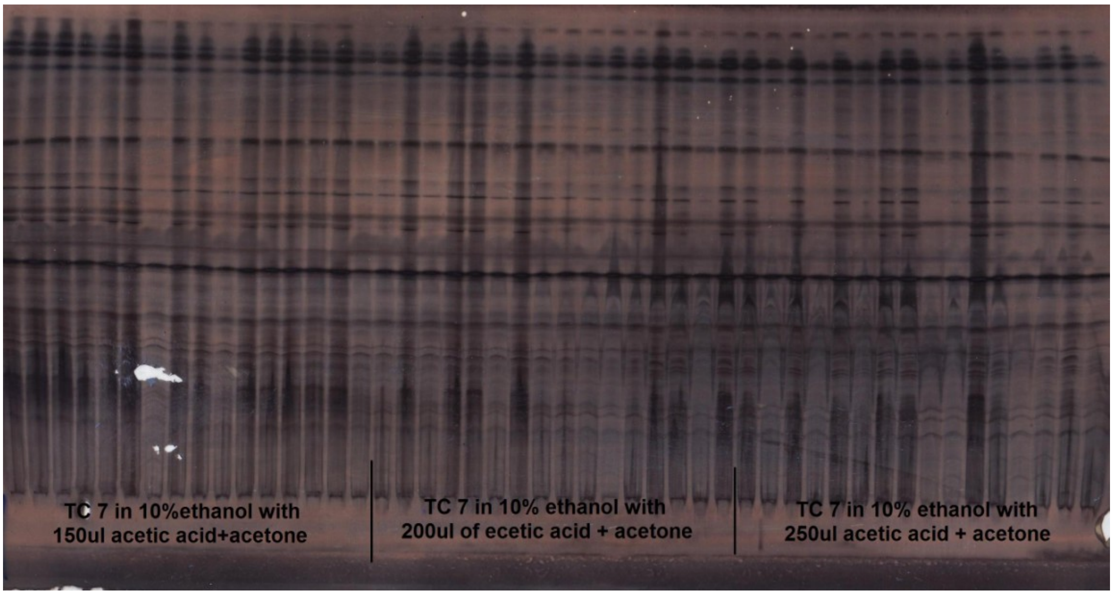
TC 7 in 10%ethanol with
150ul acetic acid+acetone
TC 7 in 10% ethanol with
200ul of ecetic acid + acetone
TC 7 in 10% ethanol with
250ul acetic acid + acetone

Figure 3. The effect of three different volumes of acetic acid-acetone mix added to the extraction.

Different volumes of the acetic acid – acetone mix were added to the 1ml of extraction solution to optimize protein gel visualization, without sacrificing or affecting any proteins in the gel. The volume of acetic acid - acetone mix that generated the best results was 150µl per extraction. At this volume (Figure 3) the visual interference disappeared. Furthermore, the solvent mix had no negative effect on the image, whereas at high volumes (200 and 250ul) this was a problem.

The combination of TC7 with 10% ethanol extracted a protein combination of both 11S globulin and 2S albumin. With the addition of 150ul glacial acetic acid + acetone to the extract, some of the oil molecules were solubilised and some phenolic compounds were trapped in the oils (Figure 4). The most important step in the extraction was the centrifugation of the extract, prior to application to the gel, because this separated the extract into two layers, with the proteins localized in the bottom layer of the extract.

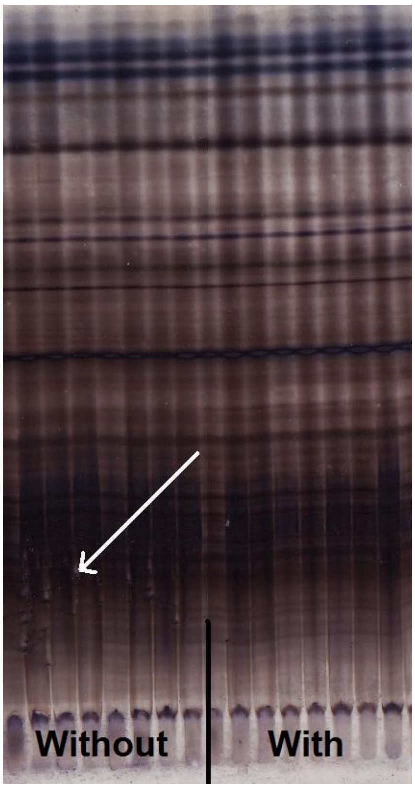

Figure 4. Protein extraction from sunflower seed: On the Left, no addition of an acetic acid + acetone mix, as indicated by the white arrow. On the Right, the addition of an acetic acid + acetone mix resulted in more clearly visualized gels, as seen in the gel on the right.

A further observation was that the total amount of seed storage proteins in each extract has a profound effect on the quality of the gel image. This varies for each sunflower inbred or hybrid, which may have small or large seeds, with high or low protein content, and has to be tested. It is essential to adjust the volume of the acetic acid-acetone mix added to the protein extract, by running a series of volumes on the new seed extraction and testing the effect in order to optimize the extraction and gel visualization of each sunflower line.

Visual interference is a global problem of electrophoresis, not just in UTLIEF but also SDS-PAGE. In a similar case of visual interference, Osset *et al*. 2005 reported that carbohydrate moieties may hinder the binding of Coomassie Brilliant Blue dyes to glycoproteins. In large commercial seed purity laboratories persistent visual interference has substantial repercussions because the quality of the gel images may be so poor that interpretation of the results are affected and samples have to be analysed again. The solution described above, to the problem of visual interference, caused by oil in sunflower extracts should significantly raise the quality of these tests and increase the efficiency of genetic purity analysis on sunflower seed protein extracts.

4.6 References

Aksyonov, I.V. 2005. Protein markers specificity of sunflower inbred lines. Helia 29: 49-54.

Aksyonov, I.V. 2005. Use of albumin markers for defining genetic purity of sunflower parent lines and hybrids. Helia 28: 43-48.

Anisimova, I.N., Gavrilova, V.A., Loskutov, A.V., Rozhkova, V.T. and Tolmachev, V.V. 2004. Polymorphism and inheritance of seed storage protein in sunflower. Russian Journal of Genetics 40: 995-1002.

Cooke, R.J. 1984. The characterization and identification of crop cultivars by electrophoresis. Electrophoresis 5: 59-72.

Fujiwara, T., Nambara, E., Yamagishi, K., Goto, D.B. and Naito, S. 2002. Storage proteins. American Society of Plant Biology. The Arabidopsis Book. http://www.aspb.org/publications/arabidopsis

Mandal, S. and Mandal, R.K. 2000. Seed storage proteins and approaches for the improvement of their nutritional quality by genetic engineering. Current Science 79: 576-589.

Osborne, T.B. 1924. The Vegetable Proteins. Longmans, Green, London.

Osset, M., Pinol, M., Fallon, M.J.M, De Lorens, R. and Cuchillo, C.M. 2005. Interference of the carbohydrate moiety in Coomassie Brilliant Blue R-250 prtein staining. Electrophoresis 10: 271-273.

Proteios B. V. 2001. Proteios protocol for gel-type 2 [unpublished protocol]. PROTEIOS B.V., the Netherlands.

Sammour, R.H. 1991. Using electrophoretic techniques in varietal identification, boisystematic analysis, phylogenetic relations and genetic resources management. Journal of Islamic Academy of Sciences 4:3, 221-226.

Shewry, P.R., Napier, J.A. and Tatham, A.S. 1995. Seed storage proteins: structure and biosynthesis. The Plant Cell 7: 945-956.

Shewry, P.R. and Halford, N.G. 2005. Cereal seed storage proteins: structure properties and role in grain utilization. Journal of Experimental Botany 53: 947-958.

Wallace, N.H. and Kriz, A.L. 1991. Nucleotide sequence of a cDNA clone corresponding to the maize Globulin-2 gene. Plant Physiology 95: 973-975.

Zhang, L.S., Le Clerc, V., Li, S. and Zhang, D. 2005. Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. Canadian Journal of Botany 83: 66-72.

CHAPTER 5: Genetic diversity analysis of sunflower using total protein and UTLIEF

5.1 Abstract

Various molecular methods have been employed to study the diversity of this globally important crop. Use of DNA-based methods is becoming commonplace in plant breeding environments, as a tool of preference to analyze and genotype plant breeding germplasm. In this study total protein profiles were generated on ultrathin layer iso-electric focusing gels (UTLIEF) (i) to assess the level of genetic diversity in elite male fertile maintainer lines (B-lines) and male fertile fertility-restoring (R-lines) sunflower lines in a proprietary breeding programme; and (ii) to compare the classification of germplasm on the basis of pedigree descriptions of individual inbred lines.

5.2 Introduction

Traditionally sunflower breeding and selection was based on morphological characters (or phenotypic characters). This approach involves the direct evaluation of plants in the field. Intellectual Property Rights on newly bred cultivars, according to the Convention of the Union Internationale pour la Protection des Obtentions Végétales (UPOV 1961), is essentially based on the ability of the parent inbred lines to display phenotypic distinctness, uniformity and stability (DUS). This is tested using phenotypic trait descriptions (Sammour 1991).

The genetic base or diversity of sunflower germplasm being used for breeding is being reduced due to the frequent use of the same genetic resources (Zhang *et al.,* 2005), resulting in a narrowing genetic base. Sunflower breeders tend to have the common objective in their breeding goals: grain yield, and abiotic and biotic stress resistance.

Sunflower is a crop that is very sensitive to G x E interactions, with the result that phenotypes of the same plant material may vary greatly according to the time and place it is grown. Furthermore, differences in the morphotype may be due to a mutation, and identical morphotypes may be created by different genes. In some cases, plants different in morphotype are genetically very similar (Aksyonov 2005). In these cases, identification of genetic variability based solely on phenotypic characteristics is not possible (Konarev 1998). Genetic distance estimation for plant registration and protection using molecular markers is becoming increasingly important for international seed companies. It is important in the scientific and commercial environment to have an economical and efficient analysis system to perform variety verification (Mitchell *et al.,* 1997; Senior *et al.,* 1998) and variety testing. Seed identity and varietal purity testing are essential components of a modern and effective agricultural production system (Nikolić 2008).

The use of molecular markers in plants has increased dramatically with the use of molecular biology techniques. With these techniques; it is now possible to identify variation at the DNA level that may not be expressed as differences in visible phenotypes. Molecular markers have many advantages (Lombard *et al.,* 2000) compared with morphological markers, resilient to environmental changes, nearly unlimited number and relative ease and rapidity of data collection. However, using these techniques need a substantial capital outlay which is not available for most of the scientists and agricultural institutions in developing countries.

Electrophoresis is an analytical tool that provides indirect access to genome probing by transcriptional variations in enzymes or other proteins, derived from the genome (Cooke 1984). There are many forms of electrophoresis separation methods available. The development of these methods has progressed from paper, cellulose acetate membranes and starch gel electrophoresis to molecular sieve, disc, SDS-PAGE and immuno-electrophoresis, and finally to iso-electric focusing including high resolution two-dimensional electrophoresis. The latest techniques enable higher resolution, sensitivity and specificity for the analysis of protein. In addition, progress in electrophoresis has been enhanced by advances in gel imaging, using silver and gold staining, autoradiography, fluorography and blotting.

The main fields of application for electrophoresis are biological and biochemical research, protein chemistry, pharmacology, forensic medicine, clinical investigations, veterinary science and food control, as well as molecular biology (Westermeier 2005).

Iso-electric focusing (IEF) is an electrophoretic method that is limited to molecules which can either be positively or negatively charged, i.e., proteins, enzymes and peptides (amphoteric molecules). Molecules are separated according to their iso-electric points (pI), in a stabilized pH gradient. The net charge of a protein is the sum of all negative and positive charges of the amino acid side chains.

The method involves casting a layer of support media, usually a polyacrylamide or agarose gel. These media contains a mixture of carrier ampholytes (low-molecular weight synthetic polyamino-polycarboxylic acids). When using a polyacrylamide gel, a low percentage gel (~4%) is used since this has a large pore size, which allows proteins to move freely under the applied electrical field. When an electric field is applied across such a gel, the carrier ampholytes arrange themselves in order of increasing pI from the anode to the cathode. Each carrier ampholyte maintains a local pH corresponding to its pI and thus a uniform pH gradient is created across the gel. If a sample of a single protein is applied to the surface of an IEF gel, then the protein will diffuse into the gel, migrate under the influence of the electric field, it will migrate until it reaches the region of the gel gradient where the pH corresponds to the protein’s specific iso-electric point. At this pH, the protein will have no net charge and will therefore become stationary in the gel. Should a protein diffuse slightly toward the anode from this point, it will gain a weak positive charge and migrate back towards the cathode, to its position of zero charge. Similarly diffusion toward the cathode results in a weak negative charge that will direct the protein back to the same position. Each protein is therefore trapped or “focused” on the gel at the pH value at which it has zero charge. Proteins are therefore separated according to their charge, and not size, as occurs with SDS-PAGE electrophoresis. In IEF it is crucial to find the correct place on the gel, i.e., point in the pH gradient, to apply each sample, because some proteins are unstable at certain pH values.

An important early step in hybrid sunflower production is to ensure that the inbred lines involved in the hybrid crosses are pure lines. UTLIEF is commonly used for the purpose of genetic purity analysis, and is the method of choice of seed producing companies, because it provides a relatively high throughput, and cost effective method that rapidly improves the quality of the seed produced (van Oers and Tamboer 2006).

The advantages of using electrophoretic markers for variety and species identification are:

a. they are rapid
b. they are relatively cheap
c. they eliminate the need to grow plants to maturity
d. they are largely unaffected by the environment.

Disadvantages include the fact that they are influenced by tissue specificity and developmental stage. This disadvantage can be overcome by evaluating seed storage proteins that are not affected by these problems.

The major components of the protein fraction of sunflower seeds are the saline solution soluble 11S globulin (helianthinin) and the water-soluble 2S albumins. Helianthinin is an oligomeric protein with a molecular mass (Mr) of approximately 305 000, made up by six spherical subunits and polypeptides with different charges. The 2S albumins consist of a heterogeneous mixture of one-chain polypeptides with an Mr of about 10, 000 – 18 000. (Anisimova *et al.,* 2004)

5.3 Materials and Methods

5.3.1 Protein Extraction

Thirty three inbred sunflower lines were screened in this study. A minimum of twenty individual seeds from each inbred line were homogenized and the seed storage proteins extracted in 1ml of extraction buffer; selecting for two different types of proteins. Firstly, the 2S albulins were selected for extraction using a buffer containing 10% ethanol. Secondly, the 11S globulin was extracted using a buffer containing 0.01M Tris – citric acid at pH 7.0 (750µl TC7 + 250µl glacial acetic acid + acetone from stock (stock solution: 3ml glacial acetic acid in 100ml acetone). The extracts were stored at -84ºC until electrophoresis was performed.

5.3.2 Electrophoresis

The extracted protein samples were applied to UTLIEF gels with a wide pI rang, using large application strips. Pre-focusing was performed on 12 x 30 PAG Type 1 and Type 2 gels (these gels were supplied by Proteios International BV). Electrophoresis was performed on a flat bed focuser (Multiphor II electrophoresis system) at a pre-cooled temperature of 10ºC. The anodal buffer consisted of 25.5mM L^{-1} aspartic acid and 24.5 mM L^{-1} glutamic acid in distilled water. The cathodal buffer consisted of 25.2 mM L^{-1} arginine, 24.6 mM L^{-1} lysine and 12% ethylenediamine in distilled water. The gels were run with one anode and one cathode each, using single direction electrophoresis. The PAG was prefocused at 200V, 30W and 12mA for 100 volt hours using a volt hour integrated electrophoresis power supply (EPS3501 – XL) (Proteios, 2001).

12 µl of each protein extract was loaded individually onto an application strip resting on the gels. Electrophoresis was performed at the following settings: gel entry run at 200 V, 30 W and 12 mA for 100 volt hours and gel focusing at 200V, 30W and 12mA for 1500 volt hours (Proteios, 2001).

After completion of protein focusing, the gels were fixed using 20% trichloroacetic acid for 15min without shaking and a further 15min with shaking. The gels were then stained using a standard Coomassie blue stain and a silver stain:

a. Fixing of the proteins in 20% TCA (trichloroacetic acid) solution,
b. Reducing the gel by washing the gels for 3 x 5min in 250ml MAD working solution (The MAD stock solution consisted of 1.5L methanol + 0.75L acetic acid. The working solution was made up with 200ml MAD stock solution, 10 mg dithiothreitol and 800 ml dH_20.). Incubate the gel in 0.1% potassium dichromate solution (prepared immediately before use) for 5min in the dark;
c. Silver stain by incubating the gel for 20min in a 0.2% silver nitrate solution (prepared immediately before use).
d. Develop the gels in 150ml of a sodium carbonate working solution (Stock Solution: 150g sodium carbonate in 1L dH_20; Working Solution: 100ml of stock solution, 400ml dH_20 and 1ml formaldehyde (37%)) for approximately 3min. The solution was changed as soon as it changed colour.
e. Stopping the development by the addition of 1% acetic acid
f. Wash and dry gels.

5.3.3 Scoring and interpretation

Gels were scored visually in a light box. The banding patterns were annotated and logged as a graphical representation of the marker data, using a programme called “GGT” (an acronym for Graphical Geno Types) (Ralph van Berloo. 2007). The data of the electro-phenograms were combined for the final analysis of the data. The various loci scored were allocated a rating based on the colour intensity of each band, ranging from 3 for a heavy dark band to 0 for the absence of a band.

5.4 Results

The results of the two different types of extracted protein are shown in Figure 1 and Figure 2, respectively.

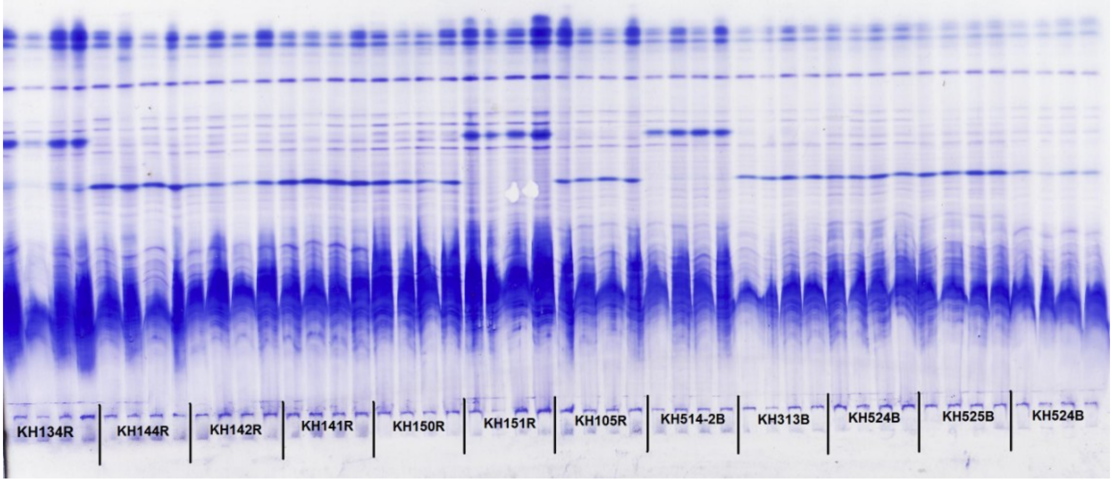

Figure 1. Gel image of the protein profile of the albumins present in 12 sunflower inbred lines

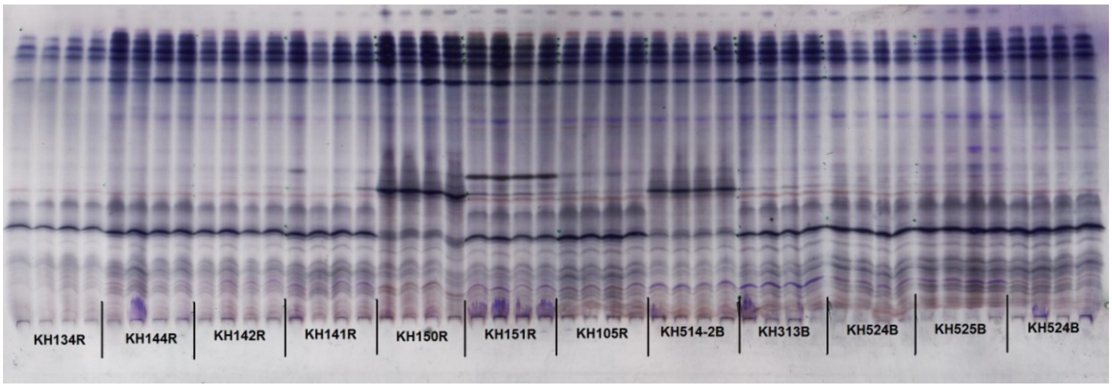

Figure 2. Gel image of the protein profile of the globulins present in 12 sunflower inbred lines

Thirty three inbred lines were genotyped using total protein markers. Two types of protein were analysed on two different gel types. A total of 68 protein bands were visualized. Genetic distance among the 33 germplasm accessions ranged from 0.03 (TF152R-TF152RHL4) to 0.145 (KH144R, KH105R, KH142-KH151R). Overall mean genetic distance was 0.426 (Figure 1).

The evolutionary history was inferred using the UPGMA method (Sneath *et al.,* 1973). The optimal tree with the sum of branch length = 3.05082634 is shown in Figure 3. The tree is drawn to scale, with branch lengths (recorded next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Phylogenetic analyses were conducted in MEGA4 (Tamura *et al.,* 2007).

This method assumes that the rate of nucleotide or amino acid substitution is the same for all evolutionary lineages. An interesting aspect of this method is that it produces a tree that mimics a species tree, with the branch lengths for two OTUs being the same after their separation. Because of the assumption of a constant rate of evolution, this method produces a rooted tree, though it is possible to remove the root for certain purposes. The algorithm for UPGMA is discussed in detail in Nei and Kumar (2000).

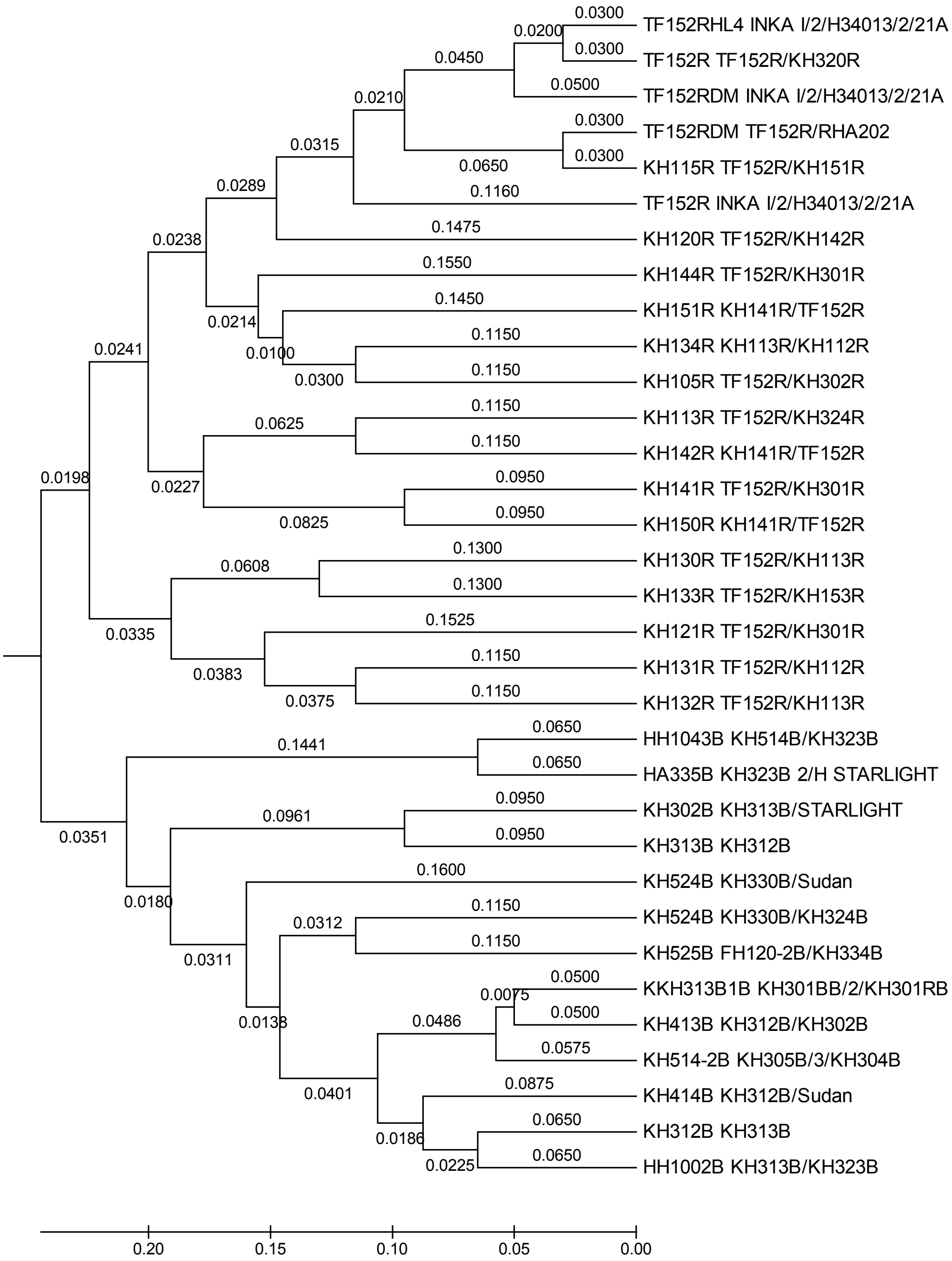

Figure 3. Evolutionary relationships of 33 inbred lines of sunflower

5.5 Discussion

UTLIEF provides for extremely high resolution images, as reflected in the two images shown above. The two different gel types show the two different types of proteins that were extracted and visualized in this study, visualized with Coomassie and silver staining.

The clustering method used was the unweighted pair group with arithmetic average clustering (UPGMA; Sneath and Sokal 1973). The dendrogram was constructed using the data derived from both types of protein extractions. This grouped the 33 genotypes in to two major clusters. The first major cluster consisted of all the R–lines (with a genetic mean of 0.382). The second major cluster consisted of the B-lines (with a genetic mean of 0.326). The smallest genetic distance values were observed between particular pairs of lines. The mean genetic distance between the isogenic TF152R lines was 0.167. In this case, the only difference between each pair of isogenic lines was supposed to be either a gene responsible for downy mildew resistance or a gene for high oleic acid. In practice there is always some residual heterogeneity between isogenic lines after the backcross procedure but this is usually a very small genetic distance. The surprisingly large genetic distance between isogenic lines tested here indicates that a relative small number of backcrosses were used to incorporate these traits into these inbred lines.

The genetic diversity study *per se* produced a strong correlation of the protein patterns with the pedigree information available for this set of inbred lines. Clear traces of the inbred lines used in previous line development could be identified in the dendrogram. Notably is the close clustering of the TF152R related lines. Some minor deviations could easily be explained by looking at other high resolution UTLIEF gel protein profiles that identified different allelic forms in some sunflower inbred lines. This could be due to the continuous improvement of the germplasm by the breeder.

When a new variety is introduced by a seed company, a reference seed lot is supplied to the molecular analysis laboratory and subsequent submissions are compared to the profile of the reference seed. Hence the laboratory staff is able to detect small differences in the protein profile of varieties when they occur. The level of heterogeneity observed in this study was low, suggesting that the cultivated sunflower inbred lines were correctly fixed. Total protein analysis performed on the same lines suggested a level of heterogeneity at the molecular level for some inbred lines.

The total protein analysis performed for genetic purity analysis on the same lines suggested a level of heterogeneity at the molecular level for some inbred lines. This can be explained by the fact that the selection of some of the sunflower inbred lines was solely based on phenotypic traits.

Of further interest would be a study to correlate genomic data derived from an SSR genetic diversity study performed on the same 33 inbred lines with the proteomic genetic diversity study presented here. This might clarify the relative advantages and disadvantages of the two approaches in terms of speed, cost and resolution of genomic versus proteomic approaches to identifying genetic diversity (and hence purity).

5.6 References

Aksyonov, I.V. 2005. Protein markers specify of sunflower inbred lines. Helia 28: 49-54.

Anisimova, I.N., Gavrilova, V.A., Loskutov, A.V., Rozhkova, V.T. and Tolmachev. V.V. 2004. Polymorphism and inheritance of seed storage protein in sunflower. Russian Journal of Genetics 40: 995-1002.

Cooke, R.J. 1984. The characterization and identification of crop cultivars by electrophoresis. Electrophoresis 5: 59-72.

Konarev, A.V. 1998. The usage of molecular markers in the investigation with plant genetic resources. Agricultural Biology 5: 3-25.

Lombard, V., Baril, C.P., Dubreuil, P., Blouet, F. and Zhang, D., 2000. Genetic relationships and fingerprinting of rapeseed cultivars by AFLP: Consequences for varietal registration. Crop Science 40: 1417–1425.

Mitchell, S.E., Kresovich, S., Jester, C.A., Hernandez, C.J. and Szewe-McFadden, A.K. 1997. Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. Crop Science. 37: 617-624.

Nikolić, Z., Vujaković, M. and Jevtić. 2008**.** Genetic purity of sunflower hybrids determined on the basis of isozymes and seed storage proteins. Helia 31, Nr. 48: 47-54.

Proteios B. V. 2001. Proteios protocol for gel-type 1 [unpublished protocol]. PROTEIOS B. V.

Proteios B. V. 2001. Proteios protocol for gel-type 2 [unpublished protocol]. PROTEIOS B. V.

Sammour, R.H. 1991. Using electrophoretic techniques in varietal identification, biosystematic analysis, phylogenetic relations and genetic resources management. Journal of Islamic Academy of Sciences 4: 221-226.

Senior, M.L., Murohy, M.M., Goodman, M.M., Stuber, C.W. 1998. Utility of SSRs for determining genetic similarities and relationships in maize using agarose gel systems. Crop Science 38: 1088-1098.

Sneath, P.H.A and Sokal, R.R. 1973. Numerical Taxonomy. Freeman, San Francisco.

Tamura, K. Dudley, J., Nei, M. and Kumar, S. 2007 MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24:1596-1599. (Publication PDF at http://www.kumarlab.net/publications)

UPOV (International Union for the protection of New Varieties of Plants). 1961. The international convention for the protection of new varieties of plants. Signed in Pars 1961. Revised in Geneva in 1972, 1978 and 1991. Available at http://www.upov.int/en/publications/conventions/.

Van Berloo, R. 2007. GGT graphical genotypes. Laboratory of Plant Breeding Wageningen University. The Netherlands. (http://www.dpw.wau.nl/pv/pub/ggt/)

van Oers, C.M. and Tamboer, J.H.A. 2006. Discriminative power of ultra thin layer gel systems in acrylamide isoelectric focusing for verification of varieties with a very narrow genetic variation [unpublished handout]. Proteios BV, the Netherlands.

Westermeier, R. 2001. Electrophoresis in Practice, Third Edition. Wiley-VCH, Weinheim, Federal Republic of Germany.

Westermeier, R. 2005. Electrophoresis in Practice. Fourth, revised and enlarged edition. Wiley-VCH, Weinheim, Federal Republic of Germany.

Zhang, L.S., Le Clerc, V., Li, S. and Zhang, D. 2005. Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. Canadian Journal of Botany 83: 66-72.

CHAPTER 6: Genetic diversity analysis of 33 sunflower inbred lines, comparing the use of genomic and proteomic analyses

6.1 Abstract

Thirty three sunflower (*Helianthus annuus* L.) inbred lines were analysed using both Simple Sequence Repeat (SSR) microsatellite markers and Ultrathin Layer Iso-electric Focusing (UTLIEF) gels of seed storage proteins. The objectives were: (i) to assess the genetic variability among these lines; and (ii), to compare the dendrogram derived from the SSRs with the dendrogram generated by the protein study. A total of 295 alleles were amplified with a set of 73 SSR markers with known mapped positions. These were utilized to determine the genetic relatedness of a group of B-line and R-line inbred lines of sunflower. In parallel, a total of 68 protein bands were visualized using protein samples of two types of seed storage proteins derived from exactly the same sunflower lines. Cluster analysis clearly differentiated between the B-lines and R-lines, identifying defined heterotic groups of this proprietary set of lines. The comparison of DNA and protein data for the application of genetic diversity studies were analysed, as well as the general comparison on the use of the two different molecules as markers. Only a limited set of phenotypic data was available for this study due to confidentiality issues. A comparison is made between the generation of DNA data vs. the generation of protein data based on the cost, speed and reliability of each type of molecule. No clear advantages were visible in the preferred use of either DNA or protein to answer the question of genetic diversity, but the strength of the combined use became clear. A combined DNA-protein analysis system is proposed to UPOV for use in plant registration and protection. Finally a breeder's tool box of molecular methods is proposed that would make a significant contribution to the speed and accuracy of a breeding programme.

6.2 Introduction

An understanding of the genetic diversity among parental lines is a major objective in plant breeding programmes aiming to develop hybrid seed. This knowledge allows the breeder to maximize genetic differences between A and B-lines, and therefore, to maximize heterosis. According to Burstin *et al.* (1994), “pedigree information provides a global estimate of the expected genetic relatedness among lines, but relies on the assumption of the absence of gametic and zygotic selection, which is often not the case”. An increasing number of molecular markers have been developed that reflect morphological and biochemical data. Previously the data sets recording genetic diversity included data based on morphological diversity (Bar-Hen *et al.* 1995), isozymes (Hamrick and Godt 1997) and storage protein profiles (Smith *et al.* 1987). These were used to assess genetic diversity among parental lines.

On a phenotypic level sunflower can be distinguished by their seed morphology:

a. seed size: short, wide, long, thin, etc.
b. seed colour: black, white or striped
c. flower morphology: the position of ray flowers
d. number of ray flowers
e. shape and colour of the ray flowers
f. head morphology: the head attitude and head size
g. leaf morphology: leaf size, shape, colour, blistering and fineness of serration
h. plant height
i. branching and type of branching.

The primary problem with using these phenotypic traits as the main criteria for genetic differences is that all of these attributes are highly sensitive to environmental changes and to the site where the plants are grown.

In contrast, the use of DNA markers has been characterized as providing "precise and reliable characterization and discrimination of genotypes", independently of the environment (Jaikishen *et al.* 2004)

Several biochemical methods, mostly electrophoresis, have also been used to estimate the genetic diversity among different plant species (Hammes *et al.* 1990). In a study on *Brassica napus*, the genetic diversity was determined from the diversity in seed storage proteins (Nasr *et al.* 2006), using SDS-PAGE electrophoresis (protein denaturing electrophoresis). SDS-PAGE electrophoresis was shown to be a powerful tool for reliable variety identification based on genetic differences in seed storage proteins.

Electrophoresis is an analytical tool that provides an indirect method for genome probing by exposing transcriptional differences reflected in enzymes or other proteins (Cooke 1984). Many methods of electrophoresis have been developed. These include electrophoresis using: paper, cellulose acetate membranes, starch gel electrophoresis, molecular sieves, discs, SDS-PAGE and immuno-electrophoresis. More recently, iso-electric focusing has been developed, including high resolution two-dimensional electrophoresis. The latest techniques enable higher resolution, sensitivity and specificity for the analysis of protein. In parallel, there have been advances in the staining of protein gels, using silver and gold stains, autoradiography, fluorography and blotting.

These genomic and proteomic techniques have been used to estimate genetic diversity, in phylogenetic reconstruction (Kaga *et al.* 1996) and plant breeding, to define the relationships between varieties, to generate linkage maps, and to identify markers linked with resistance genes against pests and diseases. However, there are pros and cons to the use of proteomic techniques versus use of genomic techniques. According to Tommasini *et al.* (2003), there is a limit to the degree of polymorphism detected by biochemical and morphological markers. Furthermore, they are often altered by the environment and the stage of plant development at sampling. In contrast, DNA-based molecular markers are unaffected by the environment and are numerous.

The detection of SSR polymorphisms has become one of the most frequently applied techniques in molecular fingerprinting (Dehmer and Friedt 1998). According to Hvarleva *et al.* (2007), SSRs are the most reliable markers for cultivar identification, genetic diversity evaluation and property rights protection. Because of their high polymorphism, random distribution, co-dominant Mendelian inheritance and their high mutation rate, they constitute the molecular markers with the highest polymorphic information content (PIC). Microsatellites are highly polymorphic, and are widely distributed throughout plant genomes. Therefore they have become one of the principle classes of DNA markers used for DNA fingerprinting, genetic mapping, and molecular breeding in crop plants (Morgate *et al.* 1993). In this study, SSRs were used to generate a phylogenetic tree for thirty three inbred lines of sunflower.

There are several advantages in the use of protein markers; genetic purity data can be generated at high speed and low cost. The genetic profile generated is the actual product of transcription and not the product of a non-functional polymorphism. According to Aksyonov (2005) "The structure of electrophoretic spectrum of seed storage proteins is not variable and it reflects the genetic makeup of the analyzed material. Therefore, electrophoretic spectrums of storage proteins may serve as reliable markers". Protein markers have an application as polygenetic markers; Singh, *et al.* (2005) describes the use of seed storage proteins to detect stable QTLs in developing drought tolerance in rice.

The goal of this study was to determine whether these two approaches, using DNA or protein markers, are comparable in their powers of discrimination, speed of throughput, ease of implementation, cost, reliability, danger to the operator, etc. The study describes the use of both SSR and UTLIEF analysis of seed storage proteins for genetic diversity analysis of the same 33 inbred sunflower lines, and looks for correlations between the two sets of results, and with established phenotypic data for the same set of sunflower inbred lines.

6.3 Materials and Methods

6.3.1 Pedigree

The following table shows the association of the 33 lines used in this study purely based on the pedigree data available.

Table 1. Summary of the pedigree of the 33 inbred lines, colours show related inbred lines.

Lines	Major Groups
TF152R 11A/INKA I/2/H34013/2/21A	R –group
TF152RHL4 11A/INKA I/2/H34013/2/21A	R –group
TF152RRM 11A/INKA I/2/H34013/2/21A	R –group
TF152R TF152R/KH320R	R –group
TF152RRM TF152R/RHA202	R –group
KH115R TF152R/KH151R	R –group
KH120R TF152R/KH142R	R –group
KH121R TF152R/KH301R	R –group
KH130R TF152R/KH113R	R –group
KH131R TF152R/KH112R	R –group
KH132R TF152R/KH113R	R –group
KH133R TF152R/KH153R	R –group
KH144R TF152R/KH301R	R –group
KH105R TF152R/KH302R	R –group
KH113R TF152R/KH324R	R –group
KH134R KH113R/KH112R	R –group
KH141R TF152R/KH301R	R –group
KH150R KH141R/TF152R	R –group
KH151R KH141R/TF152R	R –group
KH142R KH141R/TF152R	R –group
KH514-2B KH305B/3/KH304B	B-group
HA335B KH323B 2/H STARLIGHT	B-group
HH1043B KH514B/KH323B	B-group
HH1002B KH313B/KH323B	B-group
KH302B KH313B/STARLIGHT	B-group
KH313B KH312B	B-group
KH312B KH313B	B-group
KH413B KH312B/KH302B	B-group
KH414B KH312B/Sudan	B-group
KH524B KH330B/Sudan	B-group
KH524B KH330B/KH324B	B-group
KKH313B1B KH301BB/2/KH301RB	B-group
KH525B FH120-2B/KH334B	B-group

6.3.2 DNA analysis

6.3.2.1 Plant materials and isolation of DNA

DNA was isolated from 33 inbred lines. This was made up of 13 male fertile maintainer lines (B-lines) and 20 male fertile restorer lines (R-lines). Within the 33 inbred lines, some of the lines were closely related; e.g., included in the population were a parent line (TF152R), and its downy mildew resistant isogenic line (TF152RRM). There was also another parent line (TF152R) and its high oleic acid isogenic line (TF152RHL).

Genomic DNA was isolated from 7 day old seedlings, grown under controlled conditions. Five individuals per germplasm accession were harvested. Approximately 400mg of young leaf tissue was harvested, placed into a mortar and ground under liquid nitrogen. A 100mg sample of the frozen ground leaf material was weighed into an eppendorf and extracted using a Sigma Nucleic Extraction kit, according to the supplier's specifications. The concentration of the extracted DNA was determined using 0.7% TBE agarose. A working concentration of 10ng μl^{-1} was standardized for all extracted DNA.

6.3.2.2 Microsatellite genotyping

Microsatellite genotypes were produced for 33 elite inbred lines, using 73 microsatellite markers selected from a public collection (Tang *et al.* 2002; Yu *et al.* 2002). SSR genotyping primers were synthesized by Inqaba Biotech SA., and the fluorescent tails were synthesized by Applied Biosystems (Johannesburg, South Africa).

SSR genotyping was performed using an ABI3130xl sequence analyzer (from Applied Biosystems). Genotypes were identified using MapMaker 3.1 Genotyping software, also supplied by Applied Biosystems.

PCRs were performed using 12µl of reaction mixture containing 1 x PCR buffer, 2.5mM Mg^{++}, 0.2µl each of dNTPs (Bioline), 1 unit of Taq polymerase (Bioline) and 5-10ng of genomic DNA. Primers were labelled with a fluorescent dye; using a tailed primer strategy (Zhang *et al.* 2005), One tail, M13 (5'-CACGACGTTGTAAAACGAC-3'), was added to 5'-end of one of the SSR primers (forward primer) during primer synthesis. Three primers were provided for the amplification of each SSR locus: one tailed forward primer (0.05µmol), one normal reverse primer (0.25µmol) and one labelled tail (0.2µmol).

A "Touchdown" PCR was used to reduce spurious amplification. The initial denaturation step was performed at 94ºC for 2min, followed by 1 cycle at 94ºC for 30s, 63ºC for 30s and 72ºC for 45s. The annealing temperature was decreased by 1ºC per cycle in subsequent cycles until reaching a temperature of 57ºC. Products were subsequently amplified for 32 cycles at 94ºC for 30s, 57ºC for 30s, and 72ºC for 45s, with a final extension for 20min.

Amplified loci were detected by laser scanning during electrophoresis, using an ABI 3130xl Sequencer (Applied Biosystems). Samples containing 1µl of the PCR products were mixed with 8.5µl loading buffer (formamide) and 0.5µl Liz-250 internal standard (ABI). Samples were denatured at 95ºC for 5min and cooled to 4ºC, then loaded on the auto-sampler for auto injection and capillary electrophoresis. Band sizes were generated automatically, in comparison with a standard sizing ladder included in every sample prior to electrophoresis, using Genescan® and Genotyper® computer software. Band scoring was then checked manually.

6.3.2.3 Data collection and analysis

The amplification profile for each microsatellite was scored semi-automatically and evaluated. Ambiguous data were re-examined and scored manually. Bands with the same mobility were considered identical, receiving equal values. SSR markers are usually considered to reveal a single locus per primer combination. The presence of only one allele of a given microsatellite was considered a homozygous state of the allele, assuming the absence of null alleles.

The availability of marker data allows comparison of genotypes for these marker data. An overall analysis of the relatedness of all genotypes in a dataset can be performed by calculating the genetic distance between each pair of genotypes. There are several measures for estimating the genetic distance based on the marker data. For this analysis, two types of analysis were investigated:

a. the simple matching coefficient (the number of shared alleles as a proportion of all alleles)
b. the Jaccard distance or the Euclidean distance (the square root of the sum of all squared differences between alleles). The Euclidean distance is often used for quantitative data and is somewhat artificial for re-coded marker data.

Genetic distance was measured by evaluating the proportion of shared alleles per locus, polymorphic information content (PIC) and similarity values. Inbred lines were fingerprinted and therefore the selected inbreds were presumed to be homozygous for most loci. The PIC estimated the probability of observing a polymorphism between two inbred lines randomly drawn from the population of 33 lines.

A graphical representation of the molecular marker data was obtained using a programme called “GGT” (an acronym for Graphical Geno Types) (van Berloo, 2007). The data was imported into this programme, making use of commonly used marker file types that contain specified marker information. GGT data files were derived from two sources of data: A locus file, containing marker names and a raw marker scored and a (linkage) map file, specifying marker positions on a linkage map.

6.3.3 Protein

6.3.3.1 Protein Extraction:

The same 33 sunflower inbred lines were screened in this study. A minimum of twenty individual seeds from each inbred line were homogenized and the seed storage proteins extracted in 1ml of extraction buffer, selecting for two different types of proteins. Firstly, the 2S albulins were selected for extraction using a buffer containing 10% ethanol; secondly, 11S globulin was extracted using a buffer containing 0.01M Tris – citric acid at pH 7.0 (plus an acetic acid + acetone mix). The extracts were stored at -84ºC until electrophoresis was performed.

6.3.3.2 Electrophoresis:

The above extractions were applied to UTLIEF gels with a wide pI range using large application strips. Pre-focusing was performed on 12 X 30 PAG Type 1 and Type 2 gels (these gels were supplied by Proteios International, BV, the Netherlands). Electrophoresis was performed on a flat bed focuser (Multiphor II electrophoresis system from Pharmacia) at a pre-cooled temperature of 10ºC. The anodal buffer consisted of 25.5mM L^{-1} aspartic acid and 24.5mM L^{-1} glutamic acid in distilled water. The cathodal buffer consisted of 25.2mM L^{-1} arginine, 24.6mM L^{-1} lysine and 12% ethylenediamine in distilled water. The gels were run with one anode and one cathode each, in a single direction electrophoresis. The PAG was pre-focused at 200V, 30W and 12mA for 100 volt hours using a volt hour integrated electrophoresis power supply (EPS3501 – XL) (Proteios, 2001).

12 µl of each protein extraction was loaded individually onto an application strip resting on the gels. Electrophoresis was performed at the following settings: gel entry run at 200V, 30W and 12mA for 100 volt hours, and gel focusing at 200V, 30W and 12mA for 1500 volt hours (Proteios, 2001).

All electrophoresed gels were fixed in 20% trichloroacetic acid. The Type 2 gels were stained using a 0.1% Coomassie blue stain. The Type 1 gels were silver stained. To do this the gels were immersed in 20% TCA (trichloroacetic acid) solution, followed by washing of the gels for 3 x 5min in 250ml of a MAD working solution; (the MAD stock solution contained 1.5L methanol and 0.75L acetic acid; the working solution contained 200ml of the MAD stock solution, plus 10mg dithiothreitol and 800ml dH_20). After washing, the gels were incubated in 0.1% potassium dichromate solution (prepared immediately before use) for 5 min in the dark. The gels were then incubated 20min in 0.2% silver nitrate solution (prepared immediately before use). The gel was then developed in 150ml of a sodium carbonate working solution (the stock solution contained 150g sodium carbonate in 1L dH_20; the working solution contained 100ml of the stock solution, 400ml dH_20 and 1ml formaldehyde (37%)). The development stage took approximately 3min. The solution had to be changed as soon as it changed colour. The development was stopped by the addition of 1% acetic acid, after which the gels were washed, dried and annotated.

6.3.3.3 Scoring and interpretation:

Gels were scored visually in a light box. Their banding patterns were then annotated and logged as a graphical representation of the marker data, using a programme called “GGT” (an acronym for Graphical Geno Types).

6.4 Results:

6.4.1 DNA

Thirty three inbred sunflower lines were genotyped, using 73 mapped SSR markers. The markers were dispersed throughout the sunflower genome. The selected SSR markers each amplified a single locus across the 33 germplasm accessions. A total of 295 alleles were amplified using the 73 primer pairs among the 33 genotypes. The number of alleles per SSR locus varied from 2 to 9, with a mean of 4.18. The expected heterozygosity (PIC value) per locus ranged from 0.17 to 0.80, with a mean of 0.56. Genetic distance among the 33 germplasm accessions ranged from 0.02 (KH120R-KH130R) to 0.24 (KH134R-KH141R). The overall mean genetic distance was 0.574.

`

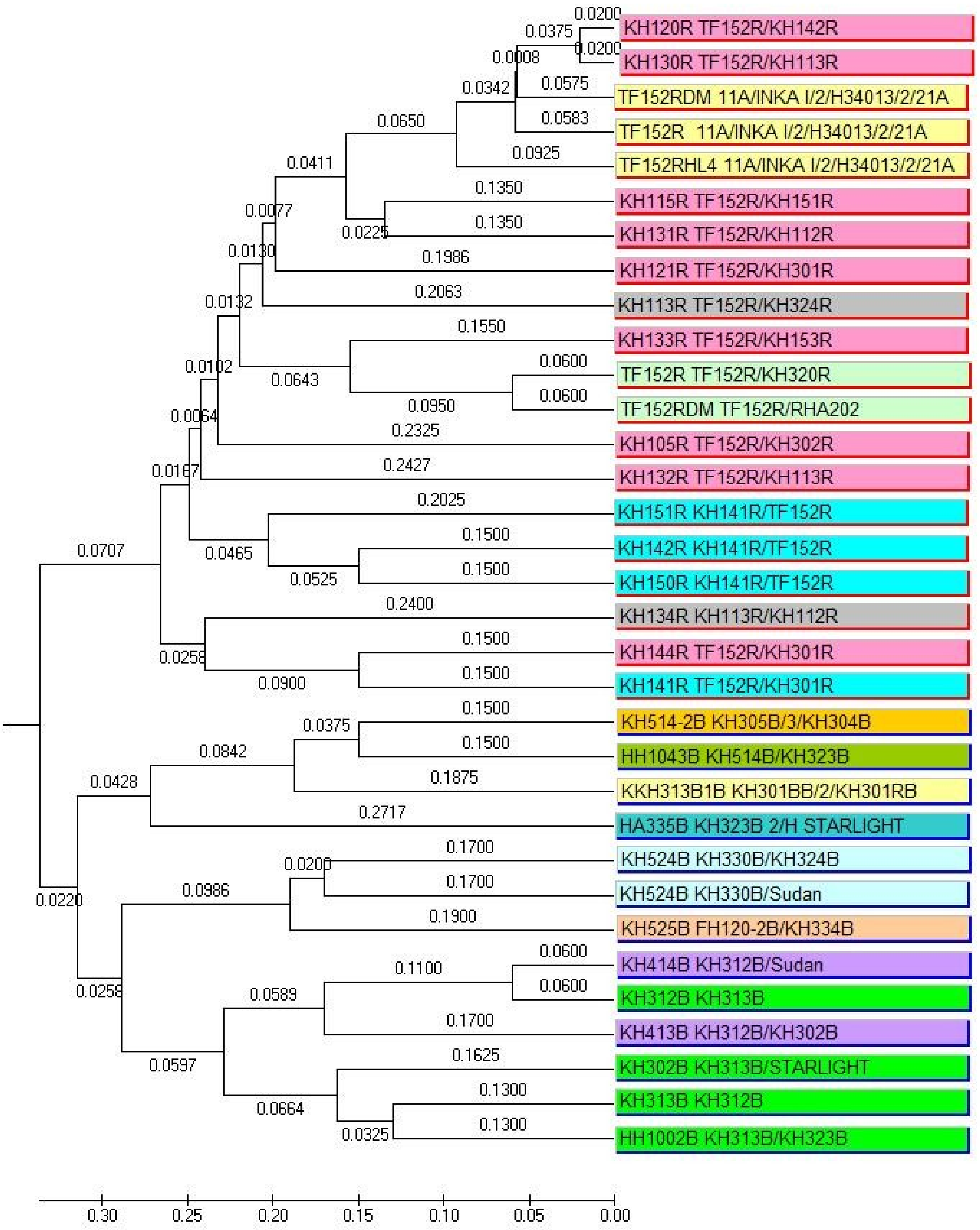

Figure 1. **Dendrogram of the 33 inbred lines of sunflower, generated from SSR data**

The phylogenetic tree was generated from the DNA data, with the evolutionary history inferred by using the UPGMA method (Sneath *et al.* 1973). The optimal tree with the sum of branch length = 6.08893399 is shown.

6.4.2 Protein

Thirty three inbred sunflower lines were analysed using protein markers . Two types of protein were analysed on two different gel types. A total of 68 protein bands were visualized. Genetic distance among the 33 germplasm accessions ranged from 0.03 (TF152R-TF152RHL4) to 0.145 (KH144R, KH105R, KH142-KH151R). Overall average is 0.426. (Figure 2).

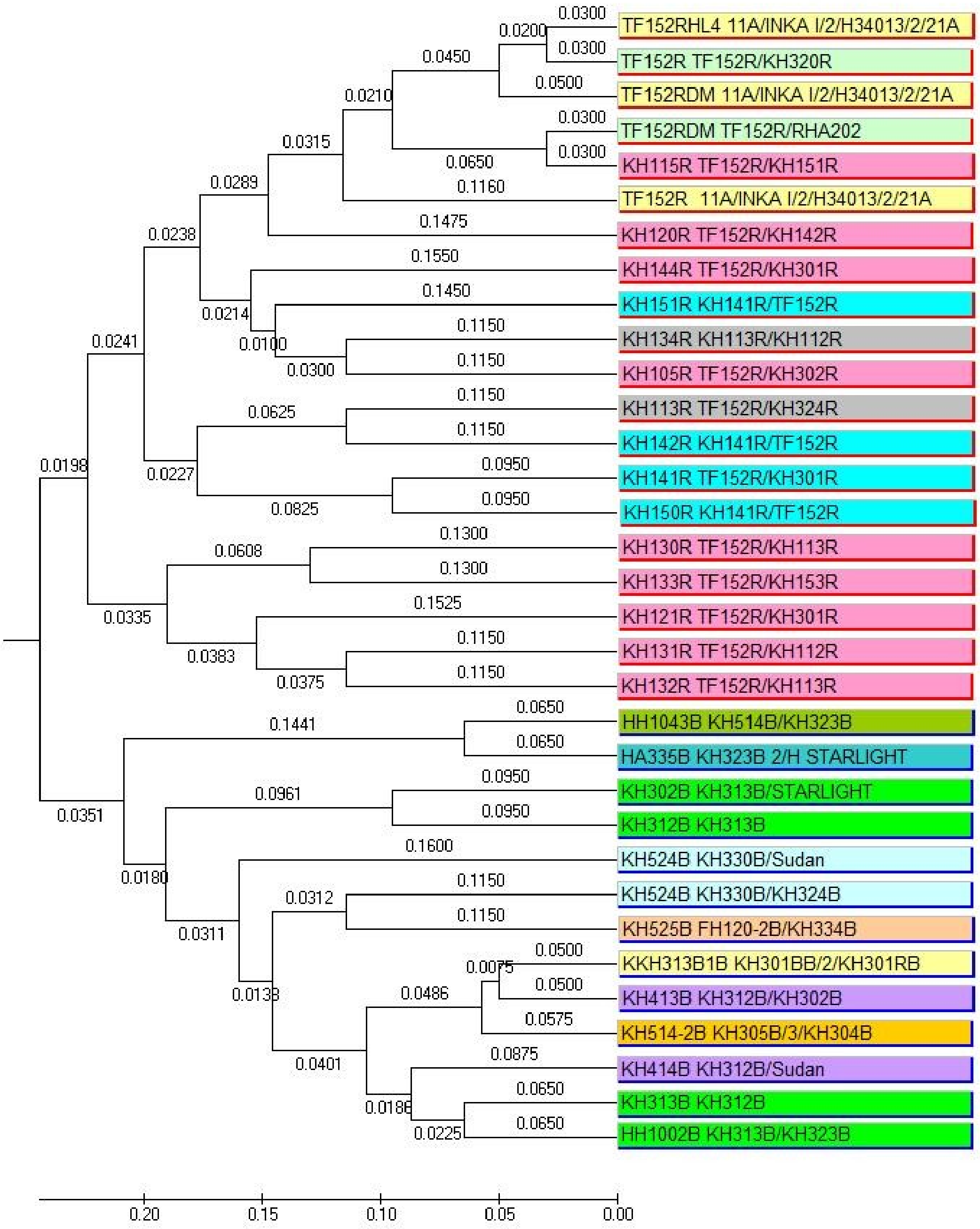

Figure 2. **Dendrogram of the 33 inbred lines of sunflower, generated from seed storage protein data**

The tree generated from the protein data used the evolutionary history inferred by using the UPGMA method (Sneath *et al.* 1973). The optimal tree with the sum of branch length = 3.05082634 is shown in Figure 3b. The trees were drawn to scale, with branch lengths (next to the branches) in the same units as those of the evolutionary distances used to infer the phylogenetic tree. Phylogenetic analyses were conducted in MEGA4 (Tamura *et al.* 2007).

6.4.3 DNA versus Protein

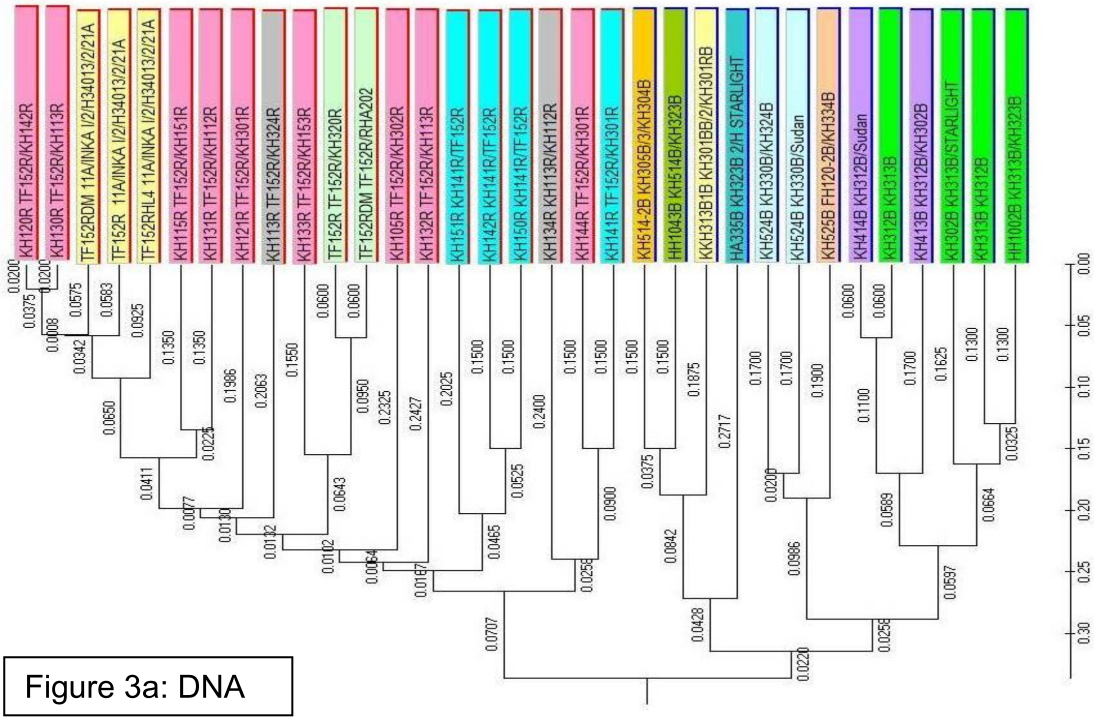

Figure 3a: DNA

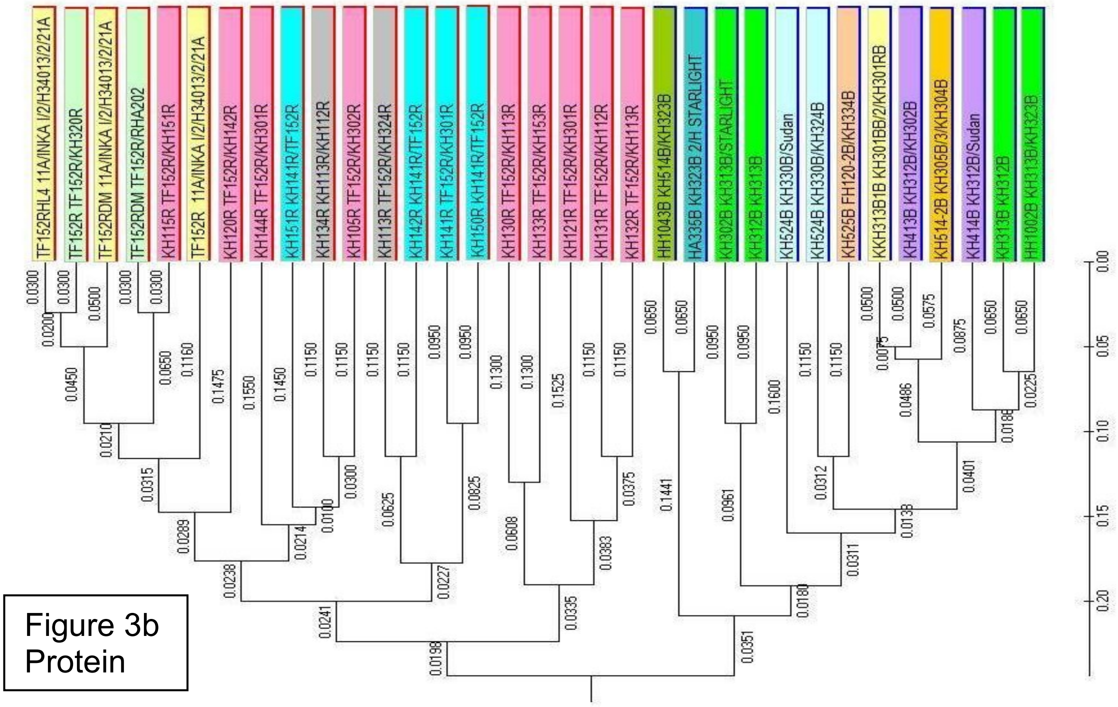

Figure 3b
Protein

Figure 3a and 3b.Phylogenic trees computed from (a) DNA analyses (at the top) and (b) Protein analyses (at the bottom).

6.4.4 Combined versus. DNA

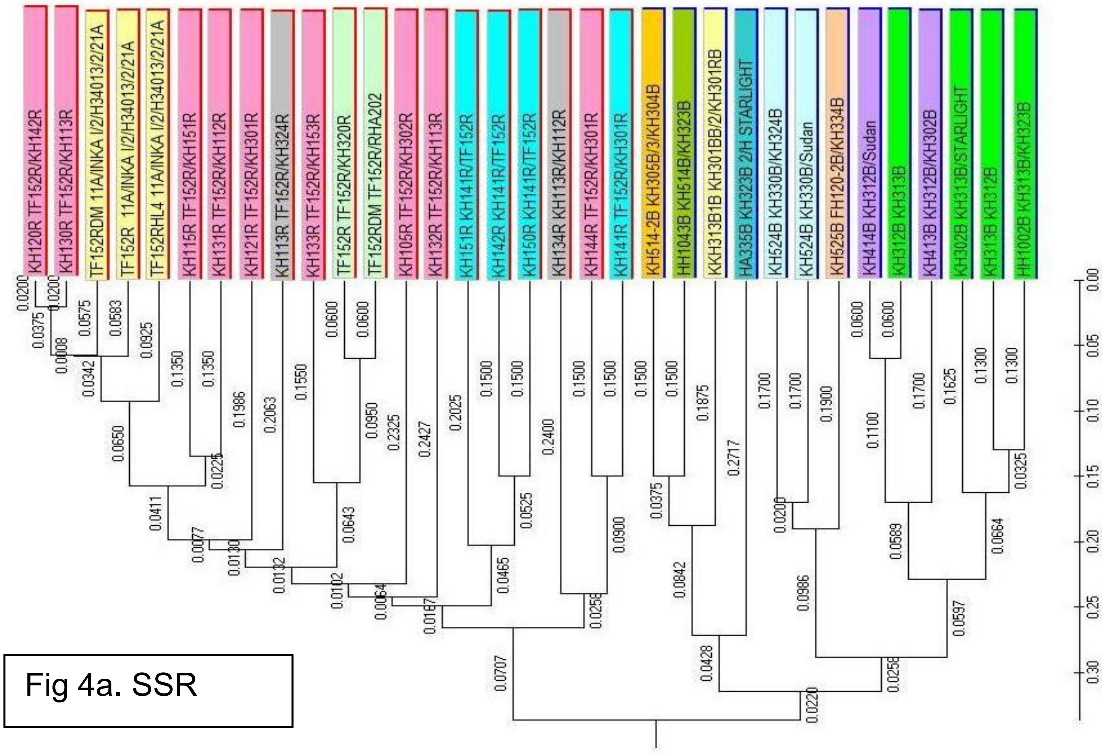

Fig 4a. SSR

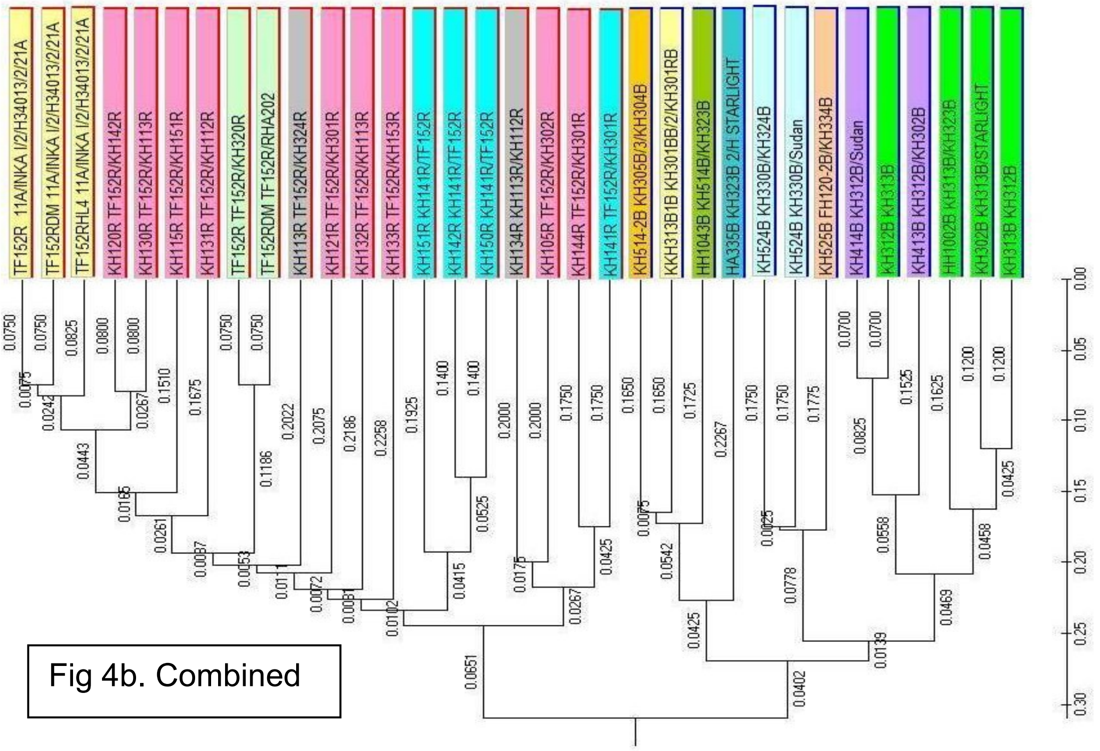

Fig 4b. Combined

Figure 4a and 4b.Phylogenetic trees computed from (a) DNA analyses (at the top) and (b) Combined DNA + Protein analyses (at the bottom).

6.4.5 Combined versus. protein

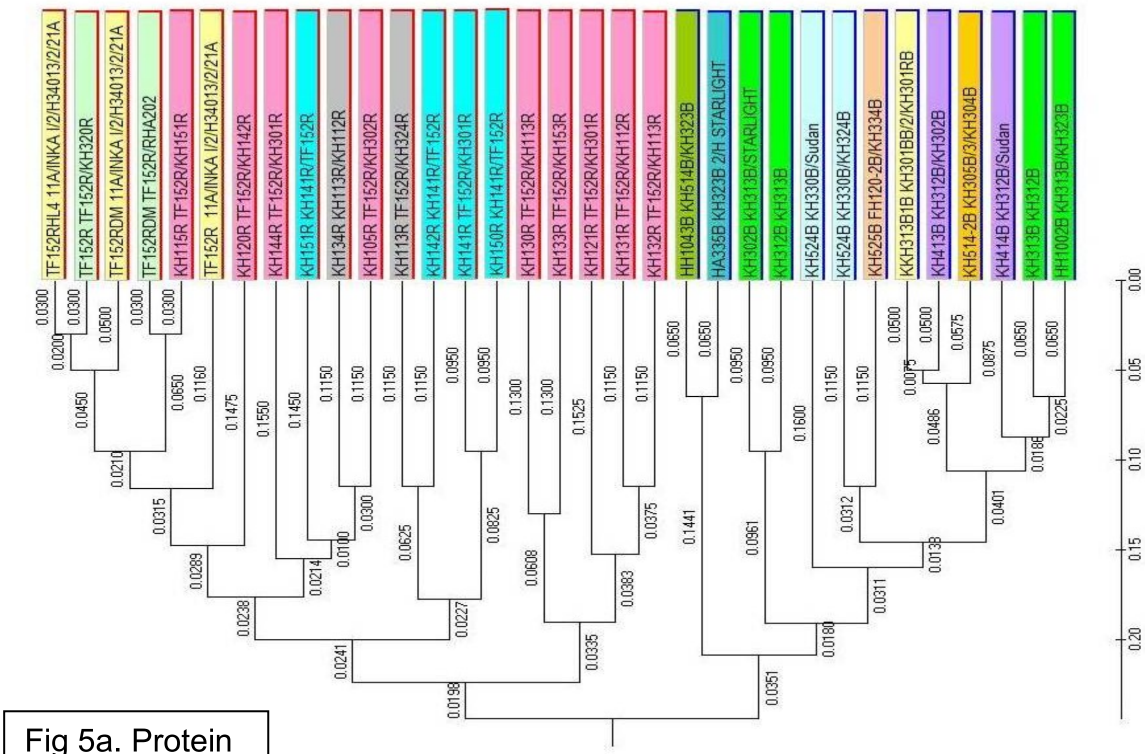

Fig 5a. Protein

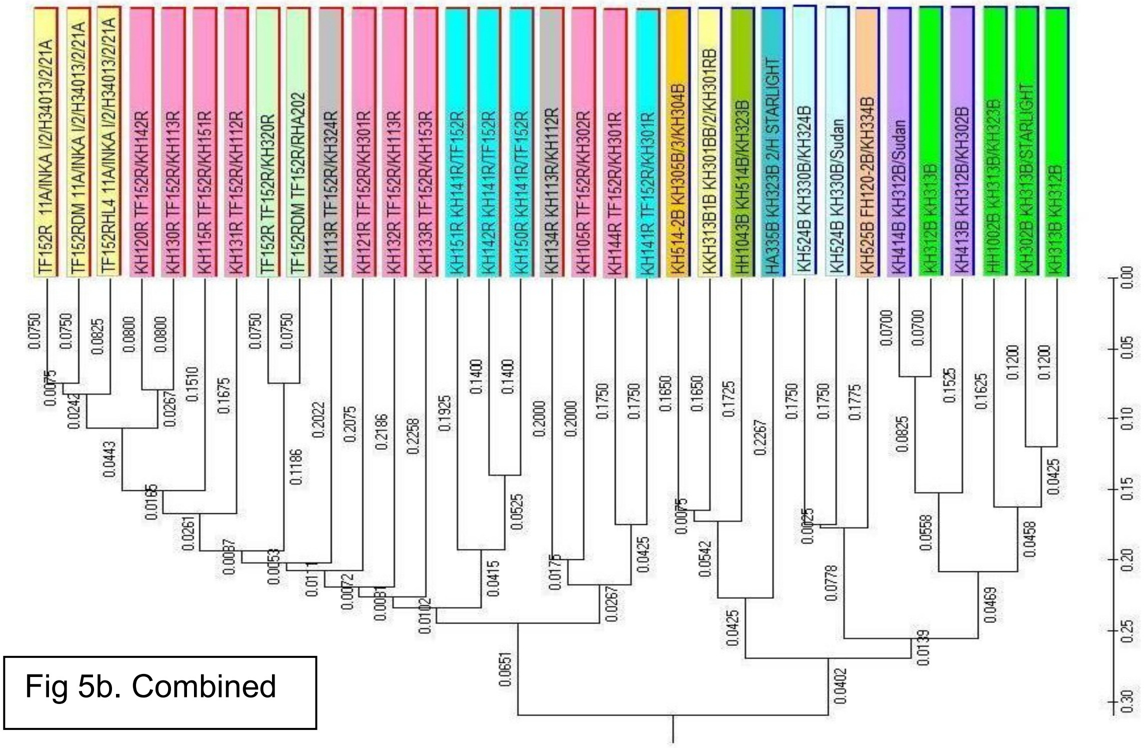

Fig 5b. Combined

Figure 5a and 5b. Phylogenetic trees computed from (a) Protein (at the top) and (b) Combined DNA + Protein analyses (at the bottom).

The data from the SSR and protein analysis was pooled to investigate the combined effect on the individual trees (Figure 4 and 5). The results from Figures 3 to 5 are compared in Table 2.

Table 2. Comparison of genetic diversity from Figures 3 to 5

Groups	DNA		Protein		Combined	
	Av. GD	Cluster	Av. GD	Cluster	Av. GD	Cluster
🟨	0.160	Inter	0.150	Weak	0.160	Strong
🟩	0.12	Strong	0.190	Weak	1.150	Strong
🟥	0.447	Weak	0.419	Weak	0.430	Weak
🟦	0.120	Weak	0.310	Weak	0.458	Inter
⬜	0.510	Weak	0.390	Weak	0.490	Weak
🟧	0.300	Inter	0.420	Weak	0.340	Weak
🟩	0.300	Inter	0.420	Weak	0.340	Weak
🟦	0.690	Weak	0.550	Weak	0.640	Weak
🟩	0.497	Weak	0.286	Weak	0.328	Inter
🟪	0.348	Inter	0.285	Weak	0.330	Inter
🟦	0.340	Inter	0.390	Weak	0.350	Inter
🟧	0.690	Weak	0.550	Weak	0.640	Weak
🟨	0.497	Weak	0.286	Weak	0.328	Weak

6.4.6 Phenotype versus Protein markers versus DNA markers

Due to confidentiality issues, data based on phenotypic information was only available for five inbred lines of sunflower. These are compared with the Protein and DNA analysis of the same five inbred lines (Figure 6a-d.)

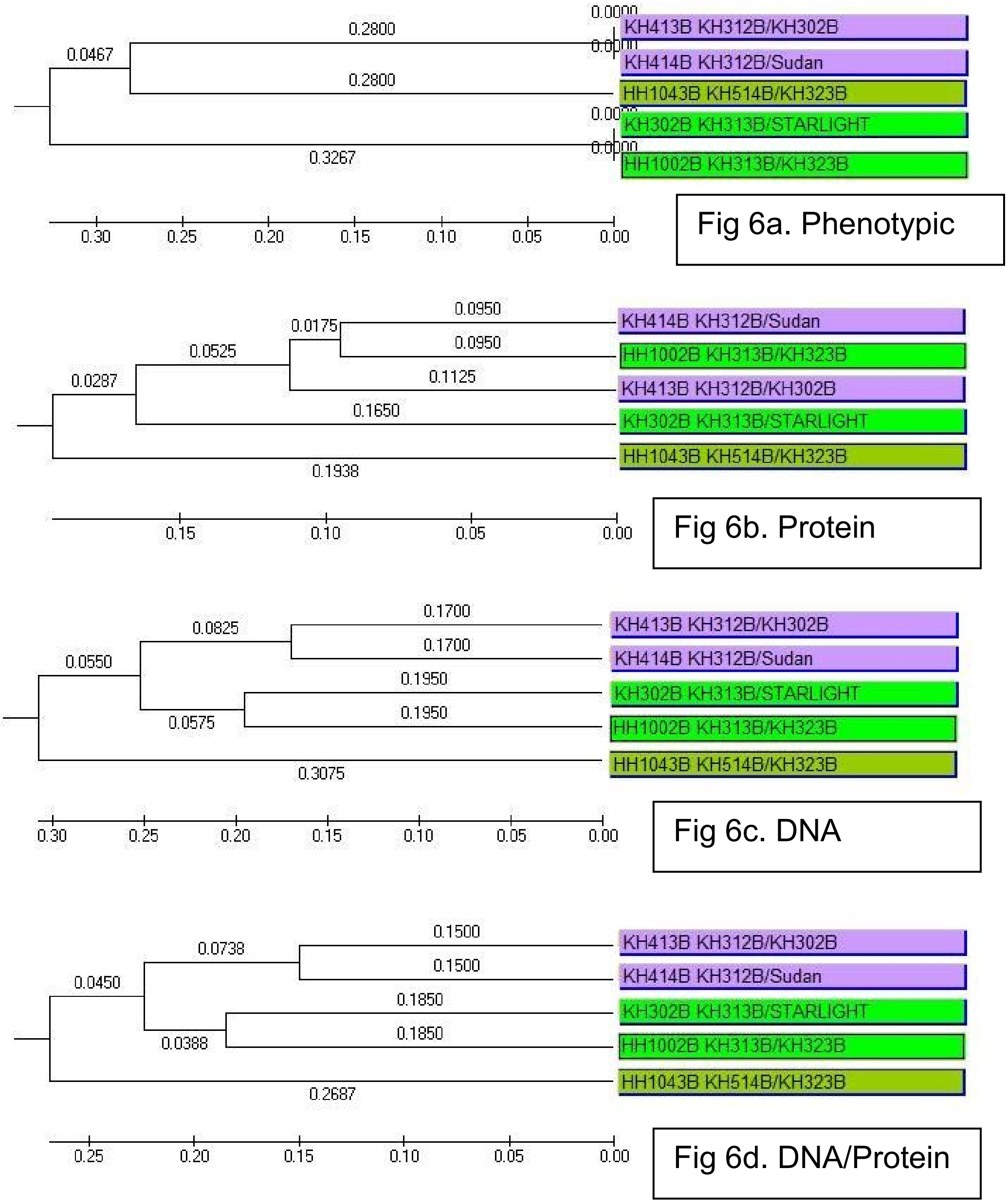

Fig 6a. Phenotypic

Fig 6b. Protein

Fig 6c. DNA

Fig 6d. DNA/Protein

Figure 6. Phylogenetic trees based on Phenotypic data (Fig 6a) versus Protein data (Fig 6b) versus DNA data (Fig. 6c) versus combined DNA-protein (Fig 6d).

6.5 Discussion

6.5.1 DNA versus Protein

The mean genetic distance computed for the DNA-based tree was 0.574, as opposed to the mean genetic distances of 0.426 calculated from the protein-based tree and 0.531 from the combined tree. The number of alleles generated from the DNA analysis was 295 from the 73 loci used to generate the data. In comparison, only 68 alleles were generated from the total protein analysis. Both trees showed a clear distinction between the two types of inbred lines that were tested: the male fertility-maintainer lines (B-lines) and male fertile fertility-restoring lines (R-lines). The genetic mean distances of the R-line cluster were 0.42; 0.382 and 0.432, on the DNA-based tree; the protein-based tree and the combined tree, respectively. In comparison, the B-line cluster, the genetic mean was 0.52; 0.326 and 0.479, on the DNA, protein and combined trees, respectively, which is much more widely spaced. This distinct difference between the genetic means of the two data sets can be explained by the fact that all male sterile sunflower inbred lines use a single source of male sterile cytoplasm, derived from a wild annual sunflower, Helianthus petiolaris (Chen et al. 2006). Hence, there is genetic convergence of all male sterile lines of sunflower.

The three isogenic lines (TF152R) were clustered together in the DNA tree with an overall average of 0.160. However, in the protein-based tree, the same three lines were not clustered together, even though they are very closely related inbreds. Their mean was 0.150, which was less than on the DNA tree. Similarly, the cluster formed by TF152R and TF152RDM (Fig. 3a, light blue circled) on the DNA-based tree had a mean of 0.120 but a bigger mean of 0.190 using the protein-based tree. One explanation for the divergence in the patterns derived from the protein versus the DNA tree was presented by Burstin *et al.* (1994), who noted that if no parental relationship existed between two lines, then the same gametic associations were not likely to be observed in the two lines, even if they were derived by one cycle of selection from the same population.

The DNA data divided the R cluster into two distinct groups. The minor group consisted of only three lines: KH134R, KH141R and KH144R. These three inbreds do not appear to be related to the main group, despite the fact that there were inbreds present in the main group with a similar genetic background to these three lines.

Of the 68 protein bands visualized, approximately 10 loci were polymorphic and capable of separating all of the male fertile maintainer B-lines from the male fertile restorer R-lines. This was made possible by the differences in allele frequency between the B and R germplasm pools. A narrow sampling of germplasm could reduce the level of polymorphism at the protein loci studied (Carrera *et al.* 2002) and the larger the numbers of molecular markers, the better the reflection of the pedigree (Hongtrakul *et al.* 1997).

The protein data divided the R cluster into three groups. The second group coalesced because of a shared pedigree of the lines in this group (KH113R, KH142R, KH141R and KH150R). The five lines in the third group also shared a similar genetic background with TF152R.

In the B cluster of both the DNA- and protein-based trees, there were three subgroups visible. In the DNA-based tree, the first group consisted of two KH514-2B related inbreds and a third inbred, KKH313B, that was expected to be unrelated (the expectation would have been to see this inbred in the third group). The second group consisted of two related KH524B inbreds and a third inbred, KH525B. The third group in this B cluster showed two distinct sub-groups: (a) consisting of three KH312B related lines; and (b) consisting of three KH313B related lines. The groupings of the B cluster on the protein tree was less clear and only two clear groups were visible: the first group consisted of two KH323B related lines; the second group can be split into five sub-groups:

a. two KH313B related lines;
b. one line, a KH524B line, that was not closely related to (c)
c. two lines, KH524B and KH525B;
d. three lines, KKH313B, KH413B and KH514B, that were expected to have no relationship with each other, based on pedigree data
e. three lines related to KH312B-KH313B: lines KH414B, KH312B and HH1002B

There have been numerous studies on genetic diversity, but few have compared the results of SSR versus protein analyses. Some authors have compared the results from RFLP markers versus enzyme analysis (McGrath and Quiros 1992; Smith and Smith 1992; Zhang *et al.* 1993). In these studies, the authors all found discrepancies between the results from the RFLP data and those from isozyme data. However, they could not determine if these differences were due to sampling bias because these two types of markers did not reveal genetic variability at the same level.

6.5.2 Combined

It is difficult to compare the use of DNA markers versus protein markers for genetic diversity analysis because each marker measures different aspects of this genetic variability. This might explain the lack of correlation between genetic diversity studies using different markers (Zeinalabendini *et al*., 2008). The two approaches produced different results. Neither is inherently superior to the other. However, the combined use of the markers could provide a far more powerful approach, by enhancing the strengths of each type of marker. The improvement of the data when the DNA and Protein data generated in this study were combined is evident when looking at Table 2, when the evaluation of the quality of the results was in comparison to known pedigree data. According to Burstin *et al.* (1994), “pedigree information provides a global estimate of the expected genetic relatedness among lines, but relies on the assumption of the absence of gametic and zygotic selection, which is often not the case”.

It is essential to take into account that the protein data was only based on two types of proteins selected for during extraction. A wider selection of proteins would have greatly improved the data. In comparison, the data generated from the use of the SSRs had an almost genome-wide coverage. A narrow sampling of germplasm could reduce the level of polymorphism at the protein loci studied (Carrera *et al.* 2002) and the larger the numbers of molecular markers (Hongtrakul *et al.* 1997), the better the reflection of the pedigree.

Table 3 and 4 compare the strengths and weaknesses of each type of marker at the level of the practical application of the marker analyses as laboratory procedures.

Table 3. **Time comparison of DNA versus Protein marker analysis (based on 96 samples)**

	DNA	Protein
Extraction time	Approx. 5 hours*	15 min
PCR time	2 hours	NA
Auto injection/Electrophoresis	4 hours	2 hours
Interpretation time	Approx. 1 hour	Approx. 20 min
Total Time	12 hours	2 hours 35 min

* Dependant on extraction method.

Table 4. **Cost comparison of DNA versus Protein marker analysis (based on 96 samples)**

	DNA	Protein
Cost of Extraction	R 1 536.00*	R 26.88
PCR cost	R 921.60 (per data point)	NA
Injection/Electrophoresis cost	R 576.00	R 680.00
Visualization (e.g. stain etc)	NA	R 260.00□
Total Cost	R 3 033.60	R 966.88

The use of SSRs gives highly reproducible and informative results. However, SSR analyses are costly and time consuming. The extraction of good quality and high yielding DNA is necessary for efficient DNA amplification. The initial costs involved in primer synthesis were high, even with the use of the tailed primer strategy. Amplification and the semi-automated analysis of the inbred lines took several months to complete because optimization is paramount in the success of any genetic diversity study, especially when using multiplexing PCR.

In contrast, the execution of SSP protein extraction from sunflower seeds, and the subsequent UTLIEF electrophoresis, used cheap, quick and robust protocols. Thousands of seed were screened daily, at a minimal cost, and the physical hands-on time was relatively short. The results of UTLIEF protein analyses are reliable and constant across multiple crops. In summary, the advent of high resolution UTLIEF gels has created the opportunity for plant breeders to undertake genetic screening of large numbers of plants on a scale that is not feasible with DNA-based techniques at present.

6.5.3 Phenotypic versus. Protein versus. DNA

The differences in the data generated from the phenotype, vs. proteins and DNA is clearly visible in Figure 6. The phenotypic analysis (based on the phenotypic characteristics listed in the Introduction) grouped KH413B KH312B/KH302B, KH414B KH312/Sudan together into a cluster (i.e., exactly the same) with HH1043 KH514B/KH323B. However, we know from pedigree information that this line is totally unrelated to the first two lines. Similarly, phenotypic grouping put KH302B KH313B/Starlight into a cluster with HH1002B KH313B/KH323B. Analysis of the protein and the DNA data show that these two lines are not related. These examples illustrated how poorly phenotypic data reflects actual genotypic variation. This study therefore created a unique opportunity to look at the efficiency of the current plant registration rules as prescribed by UPOV.

Phenotypic traits are the defined characters used for registration and plant protection by UPOV (the Union Internationale pour la Protection des Obtentions Vegetales). For protection of Plant Breeders’ Rights (PBR), parental inbred lines must be categorized in terms of distinctness, uniformity, and stability (DUS), using phenotypic trait descriptions. Due to rapid advancement in molecular techniques, the use of molecular markers in DUS testing as a complement to, or replacement of, morphological observations has become the subject of great interest in scientific studies, and consequently a topic for discussion within UPOV. However, UPOV still depends entirely upon phenotypic analyses: “Their integration (molecular markers) into DUS testing protocols still depends upon resolving of several important issues. At this point in time, all DUS testing is still based on phenotypic evaluation of the plants” (Gunjaca *et al.* 2008). With the constant improvement in molecular technology, such as is presented in this study, it is therefore proposed that UPOV should urgently implement a new approach to plant variety registrations, based primarily on molecular markers.

6.5.4 Proposal to UPOV

The need is to find a cost effective, easy-to-implement, and highly reliable system to incorporate the use of molecular markers in the plant registration process of UPOV.

Firstly, the evidence in this study makes it clear that phenotypic descriptions alone are not a strong basis for plant registration. Sunflower, in particular, is strongly affected by the environment and the season, and most hybrids produce strong G x E interactions; the phenotype of the same hybrid may vary greatly according to location and the season. These factors make the implementation of distinctness, uniformity and stability using phenotype a very difficult, and unreliable, task. If each phenotypic plant description varies from season to season because of environment, then seed companies cannot know if their registered varieties are still conforming to their documented DUS descriptions. Furthermore, breeders select for similar phenotypic traits despite using entirely different genetic material, resulting in convergent evolution of inbred lines that look similar but are genetically distinct.

Secondly, if it is accepted that molecular markers should be adopted as the basis of plant registrations, it is crucial that the technology of molecular marker use that is chosen and adopted should be quick, cheap and robust. As such, it would be accessible to virtually any plant breeding facility, either in-house or contracted out to professional laboratories.

Thirdly, the method for registration and PBR must enable a cost effective way of continuous quality control of registered plants that supersedes phenotypic evaluations. It is therefore important that the chosen molecular marker method should support the maintenance of the genetic purity of varieties as well.

An approach to the use of molecular marker data as the basis for plant breeders registration data is proposed here. It would have four main components:

- **a.** A phenotypic description because this is still useful to plant breeders;
- **b.** A genetic purity analysis based on seed proteins, using an UTLIEF analysis. The selected seed proteins would be crop specific. This is a quick, cost-effective method that can be used to determine the homogeneity of the inbred lines prior to incurring the cost of DNA genotyping. The genetic protein profile generated during this analysis could be used for future maintenance of the genetic purity of the inbred lines and varieties;
- **c.** DNA genotyping, using optimal core sets of SSRs (established for each crop), with genome-wide coverage, that can be analysed in PCR multiplex reaction for speed and cost effectiveness;
- **d.** Ongoing genetic purity analysis of registered varieties through the use of seed protein analyses, using UTLIEF.

This four step approach would solve a global problem seriously affecting seed companies and undermining the credibility of the UPOV system of plant registrations. It would provide a significant improvement to the current UPOV system based on phenotypes and the concept of DUS.

6.5.5 A "toolbox" of molecular tools for plant breeders

Molecular markers are powerful tools for plant breeders. The challenge is to generate the correct answer for each question, or to choose the most informative, cost-effective marker to apply in each breeding situation. Most major seed companies have committed themselves to using molecular tools, and many have invested millions of dollars in the development and optimization of even a single molecular technique. For plant breeders, the power of molecular technology now lies in the appropriate use of a wide range of molecular tools that are now available.

Genetic Purity

It is important to start with pure inbred material. Growing out of plants ("grow-outs") has been traditionally used to determine genetic purity. However, this is tedious, time-consuming and vulnerable to environmental changes. The most informative, cost-effective tool to evaluate genetic purity is UTLIEF of seed proteins. Within days, reliable information is available on the purity and level of inbreeding of the material tested.

Application in Breeding Programmes

Plant breeders typically use a diallel mating design to analyse for unknown traits, aiming to determine the Specific and General Combining Abilities (SGA and CGA analysis) of the parents. The diallel analysis also reveals whether the key trait is polygenic or monogenic, and additive, recessive or dominant. Once these have been determined, molecular markers can assist a plant breeder to implement this information in a practical breeding programme:

a. When breeding for polygenetic, additive traits, the use of protein markers using UTLIEF is preferable because it is fast, non-destructive and is efficient when looking for polygenetic traits.

b. When breeding for monogenetic traits (dominant or co-dominant), mapped SSRs should be the method of choice. SSRs are co-dominant markers, and they are mapped to specific chromosomes, so it is relatively easy to select for specific monogenic traits.

If genetic information is required on heterotic groups, then AFLPs should be used. AFLP are dominant markers that generate a large amount of information per primer used. Furthermore, the genetic information is random across the genome for ultimate coverage of the genome

By applying this kind of approach, most plant breeding strategies can be accelerated and enhanced by the use of appropriate proteomic and genomic tools.

6.6 References

Bar-Hen, A., Charcosset, A., Bourgoin, M. and Cuiard, J. 1995. Relationships between genetic markers and morphological traits in a maize inbred lines collection. Euphytica 84: 145-154.

Burstin, J., de Vienne, D., Dubreuil, P. and Damerval. 1994. Molecular markers and protein quantities as genetic descriptors in maize. I. Genetic diversity among 21 inbred lines. Theoretical and Applied Genetics 89: 943-950.

Carrera, A.D., Pizarro, G., Poverene, M., Feingold, S., León, A.J. and Berry, S.T. 2002. Variability among inbred lines and RFLP mapping of sunflower isozymes. Genetics and Molecular Biology 25: 65-72.

Chen, J., Hu, J., Vick, B.A. and Jan, C. C. 2006. Molecular mapping of a nuclear male-sterility gene in sunflower (*Helianthus annuus* L.) using TRAP and SSR markers. Theoretical and Applied Genetics 113: 122-127.

Cooke, R.J. 1984. The characterization and identification of crop cultivars by electrophoresis. Electrophoresis 5: 59-72.

Dehmer, K.J. and Friedt, W. 1998. Evaluation of different microsatellite motifs for the analysing genetic relationships in cultivated sunflower (*Helianthus annuus* L.). Plant Breeding 117: 45-48.

Hammes, B.D. and Richwood, M. 1990. Gel Electrophoresis of Proteins, a Practical Approach. Oxford University Press, UK.

Hamrick, J.L. and Godt, M.J.W. 1997. Allozyme diversity in cultivated crops. Crop Science 37: 26-30.

Hongtrakul, V., Huestis, G.M. and Knapp, S.J. 1997. Amplified fragment length polymorphisms as a tool for DNA fingerprinting sunflower germplasm: genetic diversity among oilseed inbred lines. Theoretical and Applied Genetics 95: 400-407.

Hvarleva, T., Bakalova, A., Chepinski, I., Hristova-Cherbadji, M., Hristov, M. and Atanasov, A. 2007. Characterization of Bulgarian sunflower cultivars and inbred lines with microsatellite markers. Biotechnology and Biotechnology Equations 21: 408-412.

Jaikishen, I., Ramesha, M.S., Rajendrakumar, P. Rao, K.S., Neeraja, C.N. Balachandran, S.M., Viraktamath, B.C., Sujatha, K. and Sundaram, R.M. Characterization of genetic diversity in hybrid rice parental lines using EST-derived and non-EST SSR markers. Rice Genetics Newsletter 23: 24-28.

Kaga. A., Tomooka, N., Egava, Y., Hosaka, K. and Kamijima, O. 1996. Species relationship in the subgenus *Ceratotropis* (genus *Vigna*) as revealed by RAPD analysis. Euphytica 88: 17-24.

McGrath, J.M., Quiros, C.F. 1992. Genetic diversity at isozyme and RFLP loci in *Brassica campestris,* as related to crop types and geographical origin. Theoretical and Applied Genetics 83: 783-790.

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

Nasr. N., Khayami, M., Hedari, R. and Jamei, R. 2006. Genetic diversity among selected varieties of *Brassica napus* (Crucifereae) based on the biochemical composition of seeds. Journal of Science (University of Tehran) 32: 37-40.

Singh, H.P., Singh. B.B and Charturvedi, G.S. 2005. Stress protein (SDS-PAGE) for MAS-breeding: Seed characteristics and vigour to detect stable QTLs using seed protein markers in developing drought tolerance in rice (O. sativa L.). Cimmyt. VI. Marker assisted selection.

Smith, J.S.C., Paszkiewicks, S., Smith, O.S. and Schaeffer, J. 1987. Electrophoretic, chromatographic and genetic techniques for identifying associations and measuring genetic diversity among corn hybrids. In Proceedings 42nd Annual Corn Sorghum Research Conference, Chicago, IL. American Seed Trade Association., Washington, DC. 187-203.

Smith, J.S.C. and Smith, O.S. 1992. Measurement of genetic diversity among maize hybrids; a comparison of isozymic, RFLP, pedigree, and heterosis data. Maydica 37: 53-60.

Sneath, P.H.A and Sokal, R.R. 1973. Numerical Taxonomy. Freeman, San Francisco.

Tamura, K. Dudley, J., Nei, M. and Kumar, S. 2007 MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596-1599. (Publication PDF at http://www.kumarlab.net/publications)

Tommasini, L., Batley, J., Arnold, G.M., Cooke, R.J., Donini, P., Lee, D., Law, J.R., Lowe, C., Moule, C., Trick, M. and Edwards, K.J. 2003. The development of multiplex simple sequence repeats (SSR) markers to compliment distinctness, uniformity and stability testing of rape (*Brassica napus* L.) varieties. Theoretical and Applied Genetics 106: 1091-1101.

Van Berloo, R. 2007. GGT graphical genotypes. Laboratory of Plant Breeding Wageningen University. The Netherlands. (http://www.dpw.wau.nl/pv/pub/ggt/)

Zeinalabedini, M., Majourhat, K., Khyam-Nekoui, M., Grigorian, V., Torchi, T., Dicenta, F. and Martínez-Gómez, P. 2008. Comparison of the use of morphological, protein and DNA markers in the genetic characterization of Iranian wild *Prunus* species. Scientia Horticulturae 116: 80-88.

Zhang, Q., Saghai Maroof, M.A., Kleihofs, A. 1993. Comparative diversity analysis of RFLPs and isozymes within and among populations of *Hordeum vulgare* ssp. *spontaneum*. Genetics 134: 909-916.

CHAPTER 7: Overview

7.1 Research Goals

A primary goal of this study was on the use of SSR analysis to generate genetic diversity data from 33 inbred lines of sunflower. This goal was successfully completed and led to the development of a unique core set of SSR markers that can be used in a novel PCR multiplex tailed strategy. This strategy proved to be significantly cheaper, faster and more labour efficient than a simplex PCR, or the traditional labelled forward primer multiplex strategy. The use of this core set of SSR primers will be of great value in sunflower breeding programmes, especially for fast genotyping of new lines and varietal verification and identification.

A second goal of this study was to use protein markers in an ultra thin layer iso-electric focusing gel (UTLIEF) analysis to generate genetic diversity data from the same 33 inbred lines of sunflower.

Whilst successful, visual interference caused by the high oil content in sunflower seed confounded the purity of seed storage proteins (SSPs) extracted from sunflower seed. This caused the loss of significant information in the protein gel analyses. Visual interference is a global problem not just in UTLIEF but also in other protein electrophoresis applications, e.g., SDS-PAGE. A parallel problem occurs where carbohydrates can interfere with electrophoresis gels. Osset *et al*. (2005) reported that carbohydrate moieties hindered the binding of Coomassie Brilliant Blue dyes to glycoproteins, affecting the evenness and reliability of gel staining.

A third goal was therefore to solve the issue of visual interference of UTLIEF gels when analysing sunflower SSPs. Adjustment of the UTLIEF protocols successfully reduced visual interference and this made a significant difference to gel interpretation and efficiency of results.

A fourth goal was to compare the phenotypic, pedigree, DNA and protein data generated from the same 33 inbred sunflower lines, for genetic diversity analysis. The outcomes were interesting and informative. This comparison was unique in that most other genetic diversity studies have only used one of these analytic tools. For example, various data sets recording genetic diversity include data based on morphological diversity (Bar-Hen *et al.* 1995), isozymes (Hamrick and Godt 1997) and storage protein profiles (Smith *et al.* 1987). It was difficult to assess from the data if the DNA or the protein gave better results because the two data sets were only compared using pedigree data. However, it was clear that the most effective analysis was to use a combination of the protein and DNA data.

According to Zhang *et al*. (2005), sunflower is strongly affected by the environment and the season, and most hybrids produce strong G x E interactions; the phenotype of the same hybrid may vary greatly according to location and the season. This has serious implication given that phenotypic traits are the only defined characters used currently for registration and plant protection by UPOV. It was clear from the results, albeit based on a very small number of lines, that there were a big differences in the results obtained from the phenotype, DNA and protein analyses. The conclusion was that the continued use of the phenotype alone for registration and PBR purposes is not viable because this data is too environmentally sensitive to be reliable.

It is therefore proposed that UPOV should alter its registration and PBR requirements away from phenotypic data alone, to including proteomic and genomic data. These are far more powerful and reliable tools to identify inbred lines and plant cultivars than morphology alone. It is suggested that the WTO UPOV protocols should adopt the following approach:

a. Phenotypic data would be retained for a general morphological description for descriptive purposes;
b. Proteomic data would be used to measure genetic purity for homogeneity and variety maintenance. Typically, these would be

"protein fingerprints", based on UTLIEF or related electrophoretic techniques;

c. Genomic data would be used to "DNA fingerprint" each cultivar, variety or breeding line for plant registration and PBR purposes. These could be based on SSR or AFLP profiles, or both.

d. Protein fingerprint data could be used for cost effective maintenance of the lines.

7.2 Implications

The implications of this study are wide and diverse.

1. The development of a core set of tailed multiplex SSR markers is a technique that provides a unique way to save cost and time to study genetic diversity in sunflower. These SSR markers for sunflower create an opportunity for large scale research projects based on the reduced costs of analysis and the greater throughput that is now possible.

2. In a commercial, high-throughput laboratory, with a key function of quality control using UTLIEF as the preferred method, visual interference between bands on gels is highly detrimental to costs, efficiency and productivity because confounded gels have to be repeated. It also reduces the level of confidence in the results of such assays because the precision of UTLIEF gels suffering from visual interference is compromised. Solving the issue of interference between protein bands has major implications for the efficiency of a high throughput system of UTLIEF analysis of high oil sunflower seed extracts for purity analysis. It also allows for a much higher level of confidence in the results derived from these analyses: every band can now be discriminated from its neighbouring band, clearly and consistently.

3. The comparison of genomic and proteomic data based on the known pedigrees of the inbred lines did not give a definitive answer as to the superiority of DNA versus protein markers or vice versa. Neither gave

a perfect match of clusters and groups of the known pedigrees of the inbred sunflower lines. However, an unexpected discovery was that the combination of DNA and protein markers gave outstanding results, and filled in gaps that existed when one or the other marker was used on its own. The match of the composite genetic distances gave a much better match with the known pedigrees. Therefore the combination of the two forms of molecular marker analysis is a far more powerful tool for plant breeders.

Creating a Molecular Marker "Toolbox" for Plant Breeders

Creating a compact "toolbox" of molecular markers would be of value to classical plant breeders, who constantly face the question of what molecular tests to use to maximize plant breeding gains. In most cases, they have little background in the molecular and biotechnology fields on which to base their judgement calls, which makes their decisions fraught. The goal of the "Molecular Marker Toolbox" below is therefore to assist plant breeders in making the right choice of tests to employ for specific objectives.

A Molecular Marker Toolbox, Version 1 (expected to evolve rapidly)

There are various techniques available. The following table list but a few general techniques and show the application, throughput capabilities when semi-automated, cost per daily throughput and the start-up cost of the equipment.

Table 1. “Toolbox”

	SSR simplex	**SSR multiplex**	**AFLP**	**UTLIEF**
Choosing the tests	Single gene, known mapped marker	Genotyping, genome wide coverage	Heterotic Grouping,	Genetic purification and genetic maintenance
Possible advantages and disadvantages.	Position known, but marker often not very close to gene	Quick and give good coverage	Position of markers unknown, very random	Quick and cost effective
Expected results	Co-dominant answer as to presence of marker	Multiple data points in short time frame	A wide general genotype, dominant marker info	Genetic protein profile showing homo-or hetero-geneity.
Robustness	Fair	Fair	Poor	Good
Daily throughput	Approx. 192 samples	Approx. 1152 samples	Approx 128 samples	Approx 1600 samples
Level of optimization required.	Limited: in getting optimal marker band (once –off)	Intense: in selecting and optimizing a core set and to multiplex (once-off)	Fair: optimal PCR conditions (once-off)	Fair: optimal extraction volume based on different seed sizes
Operating cost	R 6 067.20	R 26 238.72	R 3 328.88	R 7 735.04
Costs of equipment	R600 000.00	R600 000.00	R600 000.00	R68 000.00

7.3 Future Research

This study has created many research opportunities. Some ideas that spring to mind include:

1. The core set of SSRs could be tested across sunflower genotypes from divergent sources to determine the wider applicability of the set.
2. The suggested method of identification of a core set and the subsequent labelled tailed multiplex strategy could be tested on other crops. If it works well on many crops, then it could become a standard approach. This would make genotyping cheaper and faster for plant registrations and securing of PBR.
3. The strategy and approach to solving the visual interference on the UTLIEF gels because of the high oil content of the sunflower seed protein extracts could be applied to UTLIEF analyses of other high oil

content crops, e.g., peanuts and soybean. The exact chemical composition of the suggested extraction solution might have to be adjusted for different high oil seed crops.

4. The literature mentions that visual interference causes similar problems on SDS-PAGE gels. The strategy and approach adopted here to solve the visual interference problems for UTLIEF could be adopted to solve the problem on SDS-PAGE gels.
5. The use of conflated genomic and proteomic data to measure genetic distances could be tested on a wider range of crops. The power of the conflated analyses to discriminate between plants could be significant for plant breeders, especially in hybrid breeding programmes.
6. An obvious project would be to engage with the WTO re the UPOV conventions and rules for registration of PBR that are currently in place. Using this sunflower data set as an example, they may be persuaded to test the approach proposed above on a wide range of crops, aiming to establish a globally accepted protocol based on a combination of phenotypic, genomic and proteomic data.
7. DNA analyses, even QTL approaches, have not been successful in tracking polygenetic traits such as drought tolerance. This is logical because a trait controlled by many, additive genes, sitting on multiple chromosomes, is unlikely to be captured using genomic tools. However, using proteomics to track a polygenic trait has a much higher chance of success because the additive genes combine to generate one or a few proteins governing the trait. Therefore, another powerful application for UTLIEF could be in the study of polygenetic traits using protein markers. The technology has advanced to the extent that small quantities of critical proteins can be visualized.
8. The use of UTLIEF for genetic diversity studies and for polygenetic markers could be extended to research in other kingdoms: animals, fungi, bacteria, archaea.

7.4 References

Bar-Hen, A., Charcosset, A., Bourgoin, M. and Cuiard, J. 1995. Relationships between genetic markers and morphological traits in a maize inbred lines collection. Euphytica 84: 145-154.

Hamrick, J.L. and Godt, M.J.W. 1997. Allozyme diversity in cultivated crops. Crop Science 37: 26-30.

Osset, M., Pinol, M., Fallon, M.J.M, De Lorens, R. and Cuchillo, C.M. 2005. Interference of the carbohydrate moiety in Coomassie Brilliant Blue R-250 prtein staining. Electrophoresis 10: 271-273.

Smith, J.S.C., Paszkiewicks, S., Smith, O.S. and Schaeffer, J. 1987. Electrophoretic, chromatographic and genetic techniques for identifying associations and measuring genetic diversity among corn hybrids. In Proceedings 42nd Annual Corn Sorghum Research Conference, Chicago, IL. American Seed Trade Association., Washington, DC. 187-203.

Zhang, L.S., Le Clerc, V., Li, S. and Zhang, D. 2005. Establishment of an effective set of simple sequence repeat markers for sunflower variety identification and diversity assessment. Canadian Journal of Botany 83: 66-72.