UNIVERSITÀ DEGLI STUDI DI MILANO

Scuola di Dottorato in Scienze Biologiche e Molecolari

XXV Ciclo

Indel and SNPs of *zeins* generate unique 2D-zein patterns and molecular markers suitable for *Zea mays* genotyping and insight for seed protein quality improvement

SDD: BIO/11

Sara Piccinini

Tutor: Dr. Angelo Viotti

Coordinatore: Prof. Giovanni Dehò

Anno Accademico 2013-2014

Contents

PART I

PART II

PART I

Abstract

Maize (*Zea mays*) is an important source of proteins for human and animal nutrition. However, because of the lack of lysine and the low content in methionine and tryptophan, maize's proteins are of low quality. These deficiencies mainly result from the low levels of these essential amino acids in the zein storage proteins, which account for 50% of the total protein in mature seed. In this context, the first aim of this PhD thesis has been to develop artificial zein genes encoding for polypeptides with a higher content in lysine and methionine, and capable to be sorted and correctly accumulated into the endosperm, as occur for natural zein polypeptides.

Two strategies have been employed for maize bio-fortification. First, we exploited the natural heterogeneity among α-zein genes to create a synthetic gene, ZRK, in which six arginine residues have been substituted with lysine. Then, by combining the N-terminal methionine-rich G3 sequence and the C-terminal lysine-rich region of Histone3 and Histone4 of maize, the G3H3 and G3H4 artificial genes were created, respectively. In *vitro* and *in vivo* expression analyses of these genes showed that all synthetic proteins are synthesized and accumulated into the ER membranes of either the rabbit reticulocyte/canine membrane system or of transformed tobacco protoplasts. The second aim of this thesis has been to use the wide heterogeneity of zein gene family to obtain an *intra*-species recognition tool, or individual barcode, for inbreds and Lombard varieties discrimination. Lombard varieties and maize inbreds were analysed by 2D gel protein fractionations and DNA gel blot analyses. For each genotype the 2D and Southern blot pattern were converted into a binary code, and then into a barcode. In both the approaches, each genotype was univocally identified making zeins a valuable tool for identification of maize germplasm.

State of art

1. Cereal grains: a dietary staple

Humans are monogastric organisms and cannot synthesize essential amino acids (EAA) such as histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan and valine, which need to be supplied with diet. Although, animal protein sources like egg, milk, poultry, fish and cattle-meat are considered balanced in terms of correct ratio of EAAs (WHO/FAO/UNU report, 1985), because of their high cost, they cannot be easily afforded by the people of developing countries. Moreover, meat production has negative environmental effects, including pollution through fossil fuel usage, water and land consumption. According to a 2006 report by the United Nations Food and Agriculture Organization (FAO), our diets, which are meat based, cause more greenhouse gases carbon dioxide (CO2), methane and nitrous oxide. Greenhouse gases trap solar energy, thereby warming the earth's surface. The FAO has recently estimated that livestock accounts for about 14.5 percent of anthropogenic greenhouse gas emissions estimated as 100-year CO_2 equivalents (Gerber et al., 2013), which has significant environmental and health consequences for the planet.

According to the United Nations, 30 percent of the Earth’s land mass is now used for raising animals for food (including land used for grazing and land used to grow feed crops; FAO, 2006). For example, more than 260 million acres of U.S. forest have been cleared to create cropland to grow grain to feed farmed animals. Moreover, between watering the crops that farmed animals eat, providing drinking water for billions of animals each year, and cleaning away the filth in factory farms, the farmed animal industry places a serious strain on our water supply. Nearly half of all the water used in the United States goes to raising animals for food. In this scenario, plants clearly provide a cheaper source of dietary proteins for the poor populations as well as reducing the negative impact on the environment.

Total world cereal production is about 2 billion tonnes. Of this wheat (27%), rice (28%) and maize (34%) account for the main part (Table 1, FAOSTAT, 2012). Cereals are the prominent crops worldwide, being cultivated and harvested and last for several months of the year in both the hemispheres even in extreme latitude and cover more than 65% of dietary food for both humans and most of the other animal species. Worldwide, these cereals are subjected to a very wide range of traditional and technologically advanced processes before consumption. Thus, cereal based foods vary enormously in their structural, storage, and sensory characteristics. Cereal-based foodstuffs also vary in nutritional value owing to inherent differences in nutrient content and to changes as a result of processing, which may be beneficial or detrimental. Cereals are also the raw materials for the production of alcoholic beverages and food ingredients including starches, syrups, and protein and fibre isolates. Moreover, although humans prefer animal proteins, most animal protein is derived from the grains of cereals and, to a lesser extent, vegetable species.

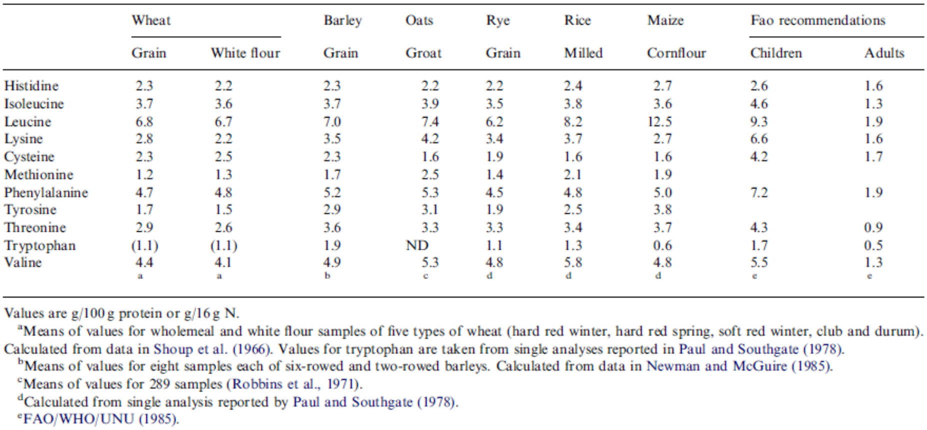
Types of cereals	Production in tonnes
Quinoa	82510
Triticale	1367102
Rye	14562055
Oats	21062972
Millet	29866016
Sorghum	57004922
Barley	132886519
Wheat	670875110
Rice, paddy	719738273
Maize	872066770

Tab. 1: World production of cereals; data are in tonnes (data from the web site of FAOSTAT)

Most of the cereal grains have an unbalanced ratio and percentage of some EAA to fulfil humans and other monogastric animals and, therefore, cereal foods need to be supplied with those scant amino acids that are part of the nine EAA of the human diet. In developed countries, plant proteins constitute only about a third of intake: e.g. the proteins intake in the USA and UK diet is 31% and 36%, respectively. Differently, plant proteins are the major source of food (about 80%) in developing countries, of which cereals predominate: two-thirds of the world's population depends on cereal or tuber-based diets (FAO, 2004). In this case, the nutritional quality, (i.e. content of essential amino acids) of the protein as well as the amount may be important.

Protein-energy malnutrition (PEM) occurs as a result of either insufficient protein intake or low value proteins; which are two key points for creating and regenerating body tissues. PEM is currently the most widespread and serious health problem of children in developing countries. A report of FAO estimated that 925 million people, more than ever, are malnourished worldwide, and malnutrition is attributable to >2.6 million child deaths every year (FAO, 2010).

Cereal proteins predominantly consist of endosperm storage proteins (see paragraph 1.1), which are low in EAAs. The first limiting EAA is generally lysine, though, among cereals the overall amount of this amino acid can importantly vary. For example, in oats and rice, the deficiency in lysine may be only marginal, while in sorghum, maize and other millets it is more pronounced. Tryptophan is also limited in maize and some millets, while threonine and methionine may also be limited in some cereals, such as rye and wheat (Table 2)



	Wheat		Barley	Oats	Rye	Rice	Maize	Fao recommendations	
	Grain	White flour	Grain	Groat	Grain	Milled	Cornflour	Children	Adults
Histidine	2.3	2.2	2.3	2.2	2.2	2.4	2.7	2.6	1.6
Isoleucine	3.7	3.6	3.7	3.9	3.5	3.8	3.6	4.6	1.3
Leucine	6.8	6.7	7.0	7.4	6.2	8.2	12.5	9.3	1.9
Lysine	2.8	2.2	3.5	4.2	3.4	3.7	2.7	6.6	1.6
Cysteine	2.3	2.5	2.3	1.6	1.9	1.6	1.6	4.2	1.7
Methionine	1.2	1.3	1.7	2.5	1.4	2.1	1.9		
Phenylalanine	4.7	4.8	5.2	5.3	4.5	4.8	5.0	7.2	1.9
Tyrosine	1.7	1.5	2.9	3.1	1.9	2.5	3.8		
Threonine	2.9	2.6	3.6	3.3	3.3	3.4	3.7	4.3	0.9
Tryptophan	(1.1)	(1.1)	1.9	ND	1.1	1.3	0.6	1.7	0.5
Valine	4.4	4.1	4.9	5.3	4.8	5.8	4.8	5.5	1.3
	a	a	b	c	d	d	d	e	e

Values are g/100 g protein or g/16 g N.

[a]Means of values for wholemeal and white flour samples of five types of wheat (hard red winter, hard red spring, soft red winter, club and durum). Calculated from data in Shoup et al. (1966). Values for tryptophan are taken from single analyses reported in Paul and Southgate (1978).

[b]Means of values for eight samples each of six-rowed and two-rowed barleys. Calculated from data in Newman and McGuire (1985).

[c]Means of values for 289 samples (Robbins et al., 1971).

[d]Calculated from single analysis reported by Paul and Southgate (1978).

[e]FAO/WHO/UNU (1985).

Tab. 2: Comparison of the contents of essential amino acids in cereal grains and flours with the FAO recommended levels for children and adults

A simple classification of cereal proteins into following four groups is based on the pioneering work of T.B. Osborne (1924), and despite subsequent modifications, this classification has retained its importance over the years:

1) Albumins: are proteins soluble in water; the group is mainly represented by enzymes;

2) Globulins: are insoluble in water but soluble in dilute salt solution; these are the major proteins in leguminous seeds;

3) Prolamins: are soluble in aqueous alcohol and are present as major proteins in maize, wheat and barley;

4) Glutelins: are insoluble in all the above solvents but soluble in dilute acids or dilute alkalies; these are the most abundant proteins in rice.

In terms of their biological function, albumins are largely known as metabolically active proteins representing enzymes, whereas the remaining three protein classes are non-enzymatic and have been called storage proteins. They are stored during seed development and used during germination by supplying the embryo with nitrogen and sulphur during seed germination (Croy et al., 1984; Boulter and Croy, 1997). However, some of the albumin polypeptides, as in peas (Croy et al., 1984), brazil nuts (Altenbach et al., 1992) and sunflowers (Krott et al., 1991) have been known to be degraded at the time of seed germination and thus, have also been ascribed the storage function. Because of the value of cereal proteins in our nutrition and industry, it is important to study how it is possible to improve their nutritional value. Accordingly, the aim of this thesis has been to explore different molecular approaches to increase the content of EAAs, in particular lysine, in the storage proteins of maize.

1.1. The zeins

Maize is one of the three most popular cereals crops (i.e. Wheat, Rice and Maize). It is grown worldwide on approximately 161 million hectares, with an annual production of 844 million metric tons (FAOSTAT, 2010). It occupies an important position in world economy and trade as a food, feed and an industrial grain crop. Several million people in developing countries consume maize as an important staple food and derive their protein and calorie requirements from it. Thus, this crop is a potential source of protein for humans and animals.

However, the poor nutritional value of maize grain is well known and the need to improve it has been recognized for a long time (Osborne and Mendel, 1914). Most proteins, in a mature maize kernel, are contained in the endosperm and the germ. Both parts differ substantially in the content and the quality of the protein. The endosperm proteins are low in quantity as well as quality, whereas the opposite is true for germ proteins. Since endosperm constitutes the bulk of the grain, it may contribute, depending on the genetic background, as much as 80% of the total kernel protein (Landry and Moreaux, 1970). Therefore, any attempt to improve maize proteins must be focused on the endosperm tissue.

The bulk of maize endosperm protein is comprised of zein fraction. Zeins are one of the best-characterized sets of storage proteins (i.e. proteins that during maturation have not enzymatic function but instead store amino acids, which are hydrolysed during germination) and are specifically expressed during seed development, acting as a reservoir of free amino acids. Zeins are about 50% of the total proteins in mature seed (Soave et al., 1981) and 62-74% of the endosperm proteins (Hamaker 1995; Landry, 2000).

Analysis of these hydrophobic alcohol-soluble proteins by SDS-polyacrylamide gel electrophoresis (SDS-PAGE), typically reveals a mixture of polypeptides ranging in size from M_r 27 kDa to M_r 10 kDa (Figure 1).

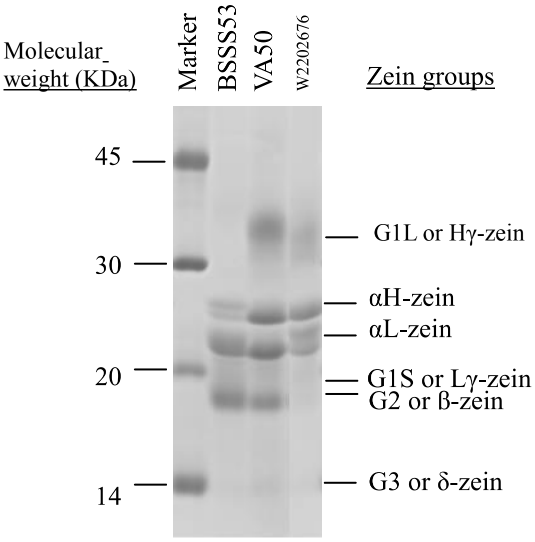


Fig. 1: One-dimensional SDS-PAGE of total zeins of tree maize genotypes

By combination of SDS-PAGE and isoelectric focusing (IEF) in a two dimensional protein separation system, zein fractions can be resolved in a number of polypeptides differing in apparent molecular weight and/or charge Figure 2 (Righetti et al., 1977).

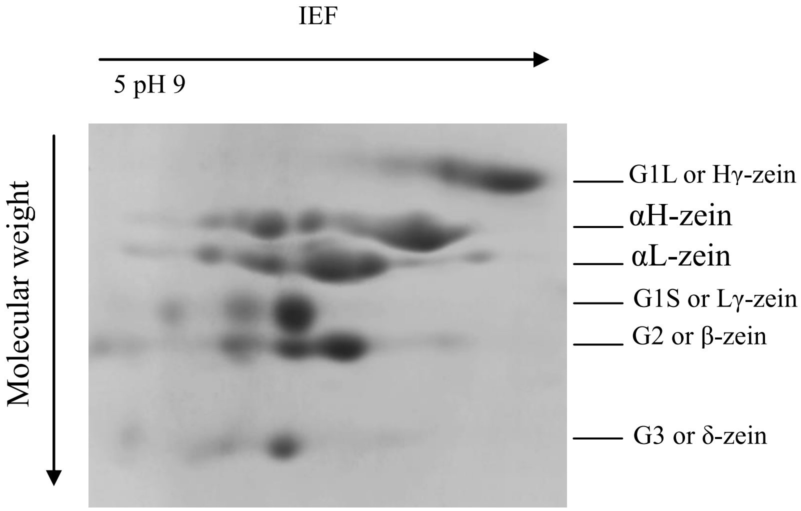


Fig.2: Two-dimensional IEF/SDS-PAGE of a total zein extract.

Based on their solubility, genetic properties, and apparent molecular masses, zeins are classified into α- (19- and 22-kDa), β- (14-kDa), γ- (16- and 27-kDa), and δ-zeins (10-kDa) and are encoded by distinct classes of structural genes. The δ-zeins are very rich in methionine, whereas the γ-zeins are abundant in cysteine; β-zein has high percentages of cysteine and methionine, while α-zeins lack lysine and are low in both sulphur aa and in tryptophan (Holding and Messing, 2013). These classes are differentially expressed during endosperm maturation, so at various developmental stages the protein bodies contain members of each family in different proportions. From an evolutionary point of view, the α- and δ-zeins arose more recently, while the γ- and β-zeins are more ancient and conserved across different subfamilies of the Poaceae (Xu and Messing, 2009). Of all these, the main protein component of maize endosperm are α-zeins, which can be further divided in four subfamilies: SF1, SF2, and SF3, coding mainly for light-class (αL) polypeptides around 19 kDa, and SF4 coding mainly for heavy-class (αH) polypeptides around 22 kDa (Viotti et al., 1985; Rubenstein and Geraghty, 1986).

1.1.1 Genomic and genic organisation of α-zein family

Alpha-zeins were among the first storage protein genes to be described. These genes have evolved from a common ancestral copy, located on the short arm of chromosome 1, to become a large multigene family (i.e. a group of genes that has descended from a common ancestor). On the basis of DNA and RNA renaturation kinetics Viotti et al., (1979) concluded that there are at least 120 genes (per haploid genome) of related sequences. Additional evidence, supporting the large size and complexity of the gene family encoding these proteins, derive from the heterogeneity observed among zeins after separation by isoelectric focusing, SDS-PAGE and extensive cDNA libraries analysis (MaizeGDB, Piccinini et al., 2014). Different haplotypes can vary, both in gene copy number and in their sequence context. The prominent feature of these haplotypes is that they can differ in the content of sequences rather than simply single nucleotide polymorphism SNPs. The latter one is the result of the tremendous transposable element activity that the maize genome has undergone after its allotetraploidization (i.e. the duplication of entire genome). That had impact not only on the expression patterns of the gene family members, with newest copies contributing the most of the mRNA pool (the older copies either accumulated premature stop codons or are truncated), but also on the mechanisms employed in their regulation, such as methylation of promoter sequences, which seems to be locus-specific. Gene bodies of storage protein genes have been shown to be undermethylated in endosperm when compared to different somatic tissues and embryo, where a common methylation pattern was reported (Bianchi and Viotti, 1988). Later a study corroborates the undermethylation in gene bodies with a CG depletion of duplicated sequences and speculates that the higher the expression of a gene is, the more CG depleted its sequence will be (Lund et al., 2003). In 2011 Miclaus and colleagues confirmed the observations and suggested that the epigenetic state of each copy is an important factor in reviving older copies from a silenced state.

All zein genes of the various classes are developmentally regulated; their expression becomes detectable c.a. 10 days after pollination and continues for c.a. 40 days. They are exclusively expressed in the endosperm of developing maize kernels and most genes of the SF4 are under the transcriptional control of the Opaque2 (O2) basic leucine zipper transcriptional activator, which binds to the O2-box in their promoter located about 300 nucleotides upstream of the translation start site.

During the development of the seed, all zein classes are synthesized on polysomes attached to a specialised part of the endoplasmic reticulum (RER) membranes (Larkins and Dalby, 1975). Secretion of the nascent protein across the reticular membrane is accompanied by the cleavage of a signal peptide sequence at the amino terminus of the native protein (Larkins et al., 1979).

The mature zeins accumulate in vesicles formed by the endoplasmic reticulum and, being insoluble continue to form large aggregates constituting the so-called protein bodies (PB_S) within the lumen of the RER (Burr and Burr, 1976;Viotti et al., 1978). Protein bodies first appear as small γ-zein accretions and later, α- and δ-zeins penetrate and appear as inclusions with the γ-zein backgrounds. Mature protein bodies of 1 to 2 µm have an even, round shape, with α-zeins confined within the core surrounded by a shell of γ- and β-zeins (Figure 3). This sequential pattern of zein protein accumulation is consistent with the temporal and spatial distribution of their mRNAs in endosperm cells (Woo et al., 2001). This model, developed from immunogold transmission electron microscopy of chemically fixed endosperm samples, did not distinguish between the locations of 19- and 22-kD α-zeins (Lending and Larkins, 1989). However, antibodies specific to 22- and 19-kD α-zeins revealed that these proteins have distinct patterns of accumulation. Whilst the 19-kD α-zein is found throughout the protein body core, the 22-kD α-zein is found only in a discrete ring at the interface between the 19-kD α-zein-rich core and the 27-kD γ-zein-rich peripheral region (Holding et al., 2007). Some evidence suggests the γ- and β-zein proteins play a role in α- and δ-zein retention in the RER (Coleman et al., 1996). When γ- and β-zeins are individually synthesized, they form protein accretions that are retained within the ER and appear to be stable over prolonged periods of time. However, when α- and δ-zeins are synthesized alone, they are not retained in the ER and become degraded. The PPPVHL, which is repeated at the N-terminus of the 27-kD γ-zein protein, could provide the mechanism for its retention in the ER (Geli et al., 1994). It was suggested that this sequence can form an amphipathic helix that interacts with the surface of the ER (Rabanal et al.,1993). Perhaps this Pro-rich sequence nucleates protein body formation, leading to interactions between γ-zeins via their conserved C-terminal regions. These proteins then could bind and retain the α- and δ-zeins, leading to their accumulation in the protein body. In 2007, Holding et al. demonstrated that also the transmembrane protein, located in the protein body ER membrane, Floury1 (FL1) is a likely coordinating factor in the assembly process.

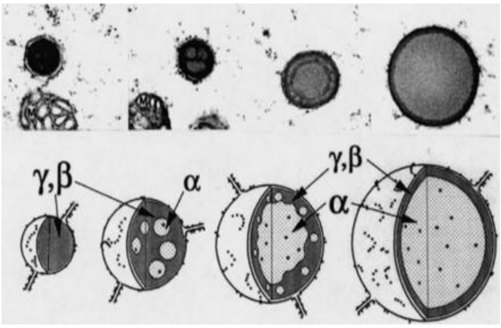


Fig.3: Developmental pattern of PB formation in maize endosperm. The most immature protein body is on the left and development is from left to right. Greek letter designations in the lower figure indicate the location of the corresponding zein classes as determined by immune-localization (Coleman et al., 1999).

Disturbance of the correct arrangement of zeins can result in irregular protein body shapes and opaque seed phenotypes (Coleman et al., 1996; Kim et al., 2004, 2008).

In 1980 Viotti et al., with in situ hybridization, discovered that α-zein genes map on different chromosomes. Later in 2001, Song et al., isolated and sequenced all 23 members of the 22-kDa α- zein gene family. Twenty-two of the 22-kD members are found in a roughly tandem array, interrupted by transposable elements and a disease response gene (CRa), on chromosome 4S forming a dense gene cluster 168,489-bp long. The twenty-third copy of the SF4 is also located on chromosome 4S at a site 20 cM closer to the centromere. Although 22-kDa α-zein gene subfamily is large, only a small number of identified genes are transcribed in detectable amounts in the endosperm (Song et al., 2001; Song and Messing, 2003). Several of the genes identified have in frame stop codons and it is postulated that many of the alpha genes are actually pseudogenes. However, we have evidence from 2D assays that some of them are expressed indicating that nucleotide transition/transversion leading to stop codons may occur during cloning and amplification of genomic fragments in the bacterial cell (Piccinini et al., 2014).

The three subfamilies of 19 kDa α-zeins map in four separate genomic regions on chromosomes 1, 4 and 7 (Song and Messing, 2002). Heavy and light zein genes are intron-less and very similar both in structure and in sequence, and contain four protein domains (Figure 4):

1. a *signal peptide* of 21 aa that contains a basic residue near the NH_2 end (lysine at the position 4) followed by a highly hydrophobic stretch at the end of which is frequently found an asparagine or threonine. The cleavage occurs after an amino acid with short side chain (alanine at position 21);
2. a *head domain* of c.a. 100 amino acids containing all or almost all the charged residues found in the mature polypeptide as well as one cysteine;
3. a *repetitive domain* composed of a variable number of block units from seven to nine. Each block is composed by c.a. 20 amino acids, which repeat a consensus sequence common to both size classes of zeins. Runs of glutamines mark transitions between blocks. The number of blocks makes the difference between the light and heavy zeins. These repetitions are predicted to have an α-helical structure and wind the protein into a rod-shaped molecule (Argos et al., 1982; Garratt et al., 1993)
4. a *short tail* piece of eight non-polar amino acids at the carboxyl end of the gene (Spena et al., 1982).

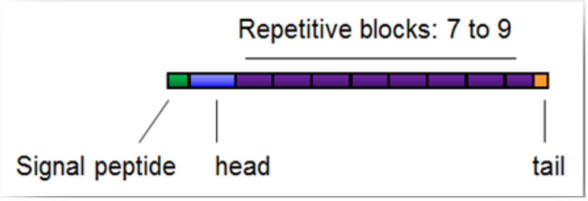


Fig. 4: Schematic representation of α-zein polypeptide.

Despite these similar structural characteristics, there is an extreme variability at the genetic level and consequently a complex situation amongst the zein proteins. The nucleotide sequence analysis reveals a greater variability in the repetitive blocks than in the other regions with insertions/deletions and substitutions resulting, at times, in drastic changes in amino acid sequences (i.e. polar amino acids into hydrophobic or vice versa). However, the three domains the signal, head and tail, are strongly conserved and characterized each subfamily (Viotti et al., 1985).

Alpha-zein proteins are characterized by a high content of proline (P), alanine (A), leucine (L) and glutamine (Q), whereas they are completely devoid of the essential amino acid lysine (K) and in some cases of tryptophan (W). The high level of this particular fraction is the sole cause of poor protein quality since all the other fractions are balanced and quite rich in lysine and tryptophan. It is thus understandable that α-zeins are the object of many researches designed to improve the nutritional value of maize.

1. 2. Approaches to obtain high-lysine maize

In the early 1960s, the analysis of the amino acid composition of seeds of the maize mutant opaque-2 (O2 is the tissue specific transcriptional activator of the most α-zein genes), revealed a double levels of lysine and tryptophan when compared with the seeds of the normal maize (Mertz et al., 1964). Characterization of *o2* mutant lines revealed that the mutation leds to changes in the relative amounts of the endosperm storage protein fractions, so that there is an increase in total lysine and tryptophan content of the mature *o2* maize seed (Tsai and Dalby 1974). However, the *o2* mutant has several drawbacks, such as reduced grain yield, soft endosperm, greater susceptibility to pests and diseases and higher moisture content at harvest time, which precluded its direct commercialization (Lambert et al., 1969). This discovery stimulated scientists to search for other mutants with increased lysine content.

Extensive breeding programs aimed to improve the o*2* seed characteristics have led to the identification of quantitative trait loci, referred to as *o2* modifiers (*Mo2s*), which convert soft into hard endosperm without losing the high lysine trait; this combination is called quality protein maize (QPM; Vasal et al., 1980). Interestingly, transcripts and proteins of 27-kDa γ-zein accumulate two- to threefold more in QPM than in normal inbreds and in unmodified *o2* mutants (Geetha et al., 1991; Holding et al., 2008). QPM has been introduced to many developing countries in South America, Africa, and Asia (Prasanna et al., 2001). However, conversion of QPM into local germplasm is a lengthy process, partly because of the technical complexity of introducing multiple *Mo2* loci, while maintaining a homozygous *o2* locus and monitoring the amino acid composition.

The second approach to developing high-lysine crops arose from the studies of the biochemistry of lysine synthesis in plants. From the 1960s until the 1980s, the biochemistry of the aspartate pathway, which leads to lysine synthesis in plant and bacteria, was extensively investigated (Azevedo and Lea, 2001). The aspartate pathway was revealed to be very complex, frequently with the end product amino acids regulating the activities of key enzymes of the pathway (Azevedo et al., 1997, 2006). Based on bacteria studies, it was envisaged that the aspartate pathway could be deregulated in plants, so that lysine and the other aspartate-derived amino acids would accumulate in tissues, including the seeds (Green and Phillips, 1974; Bright et al., 1982; Hibberd and Green 1982). By using different approaches mutants for the aspartate kinase (AK; Figure 5), which are insensitive to feedback inhibition by lysine, were selected. Although, they exhibited several-fold increases in soluble threonine, substantial lysine accumulation in the seeds was not observed (Bright et al., 1982; Rognes et al., 1983; Arruda et al., 1984). Mutants of dihydrodipicolinate synthase (DHDPS), the first aspartate pathway enzyme specifically required for lysine synthesis, were also selected and although some lysine overproduction was observed, they again failed to show major changes in lysine accumulation in the seeds (Negrutiu et al., 1984; for a review see Azevedo 2002).

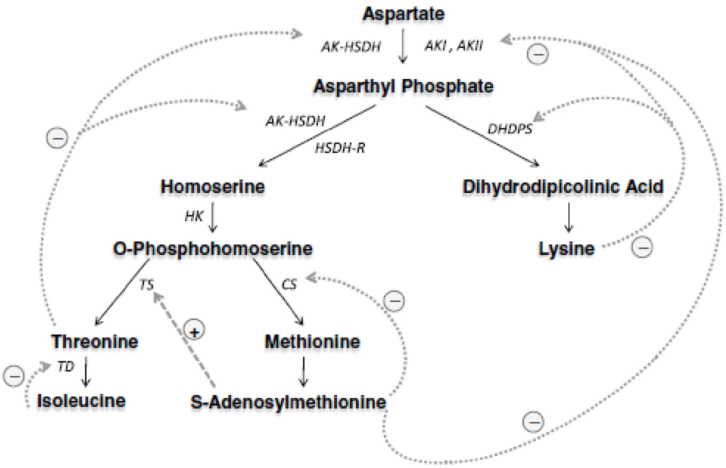


Fig. 5: The aspartate metabolic pathway of higher plants.

Only the key regulatory points are indicated. The (dotted) arrows indicate feedback inhibition or repression (-), whereas the (continuous) arrow indicates enzyme activation (+). AK I and AK II, aspartate kinase isoenzymes sensitive to lysine feedback inhibition and to lysine plus S-adenosylmethionine inhibition, AK-HSDH threonine-sensitive aspartate kinase-homoserine dehydrogenase bifunctional isoenzyme, HSDH-R threonine-resistant homoserine dehydrogenase, DHDPS lysine-sensitive dihydrodipicolinate synthase, CS cysthationine γ-synthase repressed by S-adenosylmethionine, TS threonine synthase activated by S-adenosylmethionine, and TD threonine dehydratase sensitive to isoleucine-feedback inhibition. For further detailed information about the pathway and its regulation, the following review papers should be consulted: Azevedo et al., 2006; Curien et al., 2008.

Recently through recombinant DNA technology, transgenic maize plants were engineered to knock-down zein storage protein with RNA interference (RNAi) approaches. Reduced synthesis of the lysine-poor zein proteins and compensatory increases in other proteins result in a moderate increase in total lysine content but unfortunately, some of these transgenic plants exhibit opaque phenotypes (Segal et al., 2003; Huang et al., 2006).

Another series of transgenic maize plants over-producing and over-accumulating lysine were produced using distinct strategies, including the over-expression of a bacterial DHDPS, which is less sensitive to feedback inhibition by lysine, suppression of α-zeins by antisense constructs, and both over-expression and suppression of the DHDPS and the LKR/SDH, respectively (Hournard et al., 2007; Frizzi et al., 2008; Reyes et al., 2009).

A zein gene artificially mutated to increase the number of lysine codons has been transiently expressed into maize endosperm (Torrent et al., 1997). However, most of these attempts were unsuccessful because these inherently-unstable proteins did not accumulate to sufficient levels in maize endosperm (Sun et al., 2004; Ufaz et al., 2007). Alternatively, heterologous expression of lysine-rich proteins has significantly increased lysine and total protein levels. For example, Bicar et al., (2008) introduced a construct containing a milk protein gene under the control of the seed-specific 27-kDa γ-zein promoter into the maize genome, and the transgenic lines had 29%–47% more lysine in the endosperm, although total protein content was not significantly altered. In addition, Yu et al., (2004) and Lang et al., (2004) showed that seed-specific expression of the lysine-rich cytoskeleton-associated protein genes, sb401 and SBgLR, significantly increase lysine and total protein contents in maize seeds. Furthermore, Tang et al., (2013) conducted a nutritional assessment of transgenic sb401 maize and showed that transgenic maize seeds have significantly higher levels of total protein, lysine, other amino acids, several minerals and vitamin B2 than the conventional QPM lines did. In

2014, another natural lysine-rich protein of *Gossypium hirsutum L.* (GhLRP) was introduced into the maize genome by *Agrobacterium*-mediated transformation. Transgenic maize plants accumulated GhLRP in kernels with an increase of lysine content without any damage to agronomic and quality traits of maize (Yue et al., 2014).

The testing and release of genetically modified organisms (GMOs), in particular GM plants, is tightly regulated internationally to prevent any negative effects on the environment or human health. A GM crop for cultivation could be approved by the European Union (EU), but having the most stringent regulations in the world, it is very difficult. All GMOs are subject to extensive, case-by-case, science-based food evaluation by the European Food Safety Authority (EFSA). The EFSA reports to the European Commission (EC), which then drafts proposals for granting or refusing authorisation. Each proposal is submitted to the Section on GM Food and Feed of the Standing Committee on the Food Chain and Animal Health. If accepted, it is either adopted by the EC or passed on to the Council of Agricultural Ministers. The Council has three months to reach a qualified majority for or against the proposal. If no majority is reached, the proposal is passed back to the EC, which then adopts it (Davison, 2010). Therefore, the *iter* to authorize the diffusion of GMOs is long and complicated in the EU. Italy is one of the member of the EU and has the obligation to transpose the EU directives and to comply with the regulations. Therefore, it is not possible to restrict the importation of GMO products authorized at European level or prohibit the cultivation if not scientifically supported. However, the larger part of

Italian consumers are against the use of GMOs and the diffusion of GM crops is strongly opposed by the Italian politicians.

An alternative to transgenic crop development is cisgenesis, a genetic modification of plants with cisgenes only. Although transgenesis and cisgenesis both use the same genetic modification techniques, namely the introduction of one or more genes and their promoters into a plant, cisgenesis involves only genes from the plant itself or from a close relative. Practically, these genes could also be transferred by traditional breeding techniques.

To date, the majority of established regulations on GMOs worldwide have not discriminated cisgenic from transgenic plants. This may be because until now cisgenic plants have been almost absent in applications for approval of deliberate release of transgenic plants into the environment. Only in Canada, which has a product-based regulation rather than a process-based regulation, cisgenic plants might be treated less stringently than transgenic plants. However, considering the equivalence of products resulting from cisgenesis and traditional breeding, cisgenic plants should be excluded from GMO regulations and be handled at the regulatory level like traditionally bred plants.

One example is the introduction, through the cross, of the apple scab resistance gene *Vf* from a wild apple into a cultivated one, which began as early as the 1950s (Hough et al., 1953; Schmidt and Van de Weg, 2005)

Up to now, no one has developed cisgenic transformation into maize with α-zein genes modified for increasing the content of lysine. Our approach was to consider the use of the lysine rich sequences, naturally present in maize genome, the histone 3 (H3) and histone 4 (H4). Histone genes are universally found in the genomes of eukaryotes where they encode a class of highly basic proteins, which interact each other and with the nuclear DNA to form the nucleosome: the basic structure of the chromatin. Histones can be further subdivided into subclasses according to their structural role in the chromatin fiber. The most conserved histones H3 and H4 form the central tetrameric block of the core-nuclensome; the less conserved histones H2A and H2B are added as dimers to form the final proteinaceous moiety of the nucleosome. Due to the identical role they play in the basic organization of the chromatin in all the eukaryotes, the structure of the core histones has been highly conserved through evolution (Laskey and Earnshaw, 1980; Mcghee and Felsenfeld, 1980).

We designed two synthetic genes, G3H3 and G3H4, derived from the fusion of the NH2-terminal methionine-rich G3 sequence, which contain the signal peptide for proper sorting and the C-terminal regions of H3 and H4, respectively. Co-expression (*in vitro* and/or *in vivo*) of these two genes could form stable lysine-methionine rich heterodimers as result of the known interactions of the C-terminal regions of these two histone proteins. G3H3 codes for a mature polypeptide of 22.6 kDa with a pI of 9.03 and while the other chimeric gene codes for a mature polypeptide of 21 kDa with a pI of 9.62.

Another strategy adopted was based on the identification of the natural amino acid changes along the protein sequences of the α-22kDa polypeptides. The positions of three aa, arginine and the two carboxy, were considered. Arginine and the two carboxy aa usually occur no more than three times in each polypeptide, however, in different positions among the various α-22kDa. In considering all those positions, we ended up with the ZRK gene that codes for a mature polypeptide of 24.4 kDa with a pI of 8.68. It is a α-zein sequence, with the signal peptide, in which six codons of arginine were modified in lysine codons. Moreover, four codons for the carboxy aa were introduced in proper positions to balance the positive charges introduced by lysines, which would have drastically altered the biochemical features of the α-22kDa. In all three synthetic genes the content of lysine is increased from 0 % to 3%, which is more than FAO recommendations for an equilibrate diet.

2. DNA barcoding for plants

In 2003 Paul Hebert and colleagues, a team of evolutionary biologist from the University of Guelph in Canada, proposed a way to distinguish any animal species from any other by using a short sequence of DNA (Hebert et al., 2003a). In that study, the authors explored the use of the 5'-end portion (600bp) of the Cytochrome Oxidase subunit I (COI), a mitochondrial gene, as a possible method of global bio-identification for animals. The COI gene had two important advantages for the intent of animal genotyping. First, the universal primers for the COI genes were available, thus, enabling recovery of the 5' end from almost all representatives' animal phyla (Folmer et al., 1994; Zhang and Hewitt, 1997). Second, the third-position nucleotide, in the COI sequence, shows a high incidence of base substitutions (Knowlton and Weigt, 1998). Accordingly, Hebert and colleagues (2003b) by using the COI gene system profiled 200 closely allied species and subsequently assigned 150 newly analysed individuals to species. Hebert called it an organism's "barcode."

This was not the first time that DNA barcoding had been proposed as a concept. The use of short DNA sequences to discriminate microbial species was proposed as early as 1982 (Nanney 1982) and successfully used on a variety of taxa from nematodes to elephants and even that most famous of extinct species: the dodo (Eggert et al., 2002; Floyd et al., 2002; Shapiro et al., 2002).

However, the mitochondrial gene is insufficient to discern some closely related species where mitochondrial DNA changes are too slow or too fast since species-level identification is achieved when interspecific variation exceeds intraspecific variation (often referred to as the barcode gap). For example, in 2002, Sharer and colleagues demonstrated that mitochondrial DNA of the class Anthozoa (sea anemones, corals and sea pens) has slow evolutionary rate and low sequence diversity within species, limiting its use in DNA barcoding. On the contrary, in amphibians, the most taxonomically diverse vertebrate groups, there have been consistent problems with using COI as a barcode because of the high variability of the COI priming sites (Vences et al., 2005).

To date the best DNA source for plant barcoding has been the plastid genome. This genome shares many of the desirable attributes of animal mitochondrial DNA for barcoding, such as conserved gene order and high copy number in each cell enabling easy retrieval of DNA. One problem with plastid DNA, however, is its generally slow rate of evolution, and the challenge has been to find a plastid region that is sufficiently variable for DNA barcoding. To facilitate and formalise the selection of a plant barcode, in 2004 the Consortium for the Barcode of Life (CBOL) was established. This is an international initiative devoted to developing DNA barcoding as a global standard for the identification of biological species through working groups, networks, workshops, conferences, outreach, and training. CBOL (CBOL, 2009) evaluated seven chloroplast genomic regions across the plant kingdom, four are portions of coding genes (matK, rbcL, rpoB, and rpoC1), and 3 are noncoding spacers (atpF–atpH, trnH–psbA, and psbK–psbI) in three divergent groups of land plants (angiosperm, gymnosperm, and liverwort). The criteria used for that screening are three: *Universality,* which loci can be routinely sequenced across the land plants; *Sequence quality and coverage*, which loci are most amenable to the production of bidirectional sequences with few or no ambiguous base calls; *Discrimination,* which loci enable most species to be distinguished.

Based on these criteria, 4 of the candidate loci were excluded. Both rpoC1 and rpoB performed well in terms of universality and/or sequence quality, but had low discriminatory power; atpF–atpH showed relatively modest discriminatory power and intermediate sequence quality and universality; whereas psbK–psbI showed high levels of discriminatory power, but lower sequence quality and universality. However, choosing a plant barcode from the 3 remaining candidate loci was more difficult.

The matK is one of the most rapidly evolving genes in the chloroplast genome. It encodes the enzyme maturase, which is involved in the splicing of type-II introns from RNA transcript. The matK barcode region consists of a ca. 841 bp region at the centre of the gene, located between bp 205–1046 (including primer sites) in the complete *A. thaliana* plastid genome sequence. Unfortunately, matK can be difficult to amplify using existing PCR primer sets.

The rbcL barcode consists of a 599 bp region of *Arabidopsis thaliana* plastid genome. The barcode region of rbcL is easy to amplify, sequence, and align in most land plants and provides a useful backbone to the barcode dataset, despite it having only modest discriminatory power.

The inter-genic region trnH–psbA (450 bp) is among the most variable regions in the angiosperm chloroplast genome. Moreover, it shows both good amplification across land plants with a single pair of primers and high levels of species discrimination. However, problems obtaining high quality bidirectional sequences are the primary limitation for this locus.

Individually, trnH–psbA, rbcL, and matK possess attributes that are highly desirable in a plant DNA barcoding system, although none of the 3 loci fits all 3 criteria perfectly. These results demonstrated that no single locus has high levels of universality and resolvability so synergistic combination of loci is required (Hollingsworth et al., 2009).

Later in 2011, Li D.Z. and colleagues proposed another locus for supplementary plant barcoding, the internal transcribed spacers of nuclear ribosomal DNA (nrDNA ITS). The universality of this marker (c.a. 76.5%) was lower than that of the plastidial markers (c.a. 87-93%) but it offered a significant increase in discriminatory power. However, there are three major potential problems:

1. Fungal contamination: the primers used for amplification and sequencing of nrITS in plants and fungi are similar enough such that fungal DNA is often inadvertently amplified from plant samples.
2. Paralogous gene copies: the nrDNA ITS region is present in multiple copies within each cell. These copies generally evolve in a concerted fashion, leading to a single detectable sequence per plant. However, in some plant groups, divergent copies co-occur within individuals. This can lead to messy sequences (attributable to the presence of multiple different variants being simultaneously sequenced) or, worse, inadvertent differential sequencing of different variants among samples. This process can lead to members of the same species being given different identifications depending on which variant was sequenced.

3. Recovery: The main limitation for nrDNA ITS is that it is sometimes only difficult to amplify and sequence (Kress et al., 2005). If obtaining full ITS is difficult, one can amplify up just half of the region, just ITS2 (Chen et al., 2010). This partial region is often much easier to amplify and sequence than the entire region, but can still provide appreciable gains in discriminatory power.

So up to now, after several broad screenings of gene regions in the plant genome, one nuclear (ITS) and three plastid (rbcL, matK, and trnH-psbA) gene regions have become the standard barcode of choice in most investigations for plants because they meet the criteria above mentioned.

2.1 DNA barcoding methods

The process of DNA barcoding entails two basic steps: i) building the barcode library of known species and ii) matching, or assigning the barcode sequence of the unknown sample to the barcode library for identification. For this purpose in 2010 The International Barcode of Life project was activated. Its mission is the maintenance of the barcode reference library BOLD (Barcode of Life Data System) as an international platform which supports the collection of DNA barcodes for all living species.

One of the first steps in DNA barcoding of plants is the collection of data, i.e. the complete sequences of molecular markers used from the specimen. DNA is extracted from the sample and amplified with well-developed universal primer sets for the barcode region. After the corresponding query sequences have been collected, the next step is sequence analysis and construction of phylogenetic trees. Sequence analysis basically involves the alignment of query sequences with the reference dataset of sequences. Alignment algorithms are methods that extract meaning out of raw sequences by finding regions/patterns of similarity between them. Multiple alignment step (MSA) is critical because if a bad alignment is obtained, the resulting phylogenetic tree will be incorrect leading to misidentification of the taxa, which is the primary goal of DNA barcoding. Especially, in the case of rRNA and non-protein coding markers, the high frequency of indels makes sequence alignment a critical step. Some of the MSA programs are ClustalW (Thompson et al., 1994), T-Coffe (Notredame et al., 2000), and MUSCLE (Edgar, 2004). Successful DNA barcoding requires correct identification of the query sequence from the reference database. Most conventional softwares make DNA barcoding a process based on sequence alignments, either global or local. Global alignment requires the molecular markers, e.g. COI (Saccone et al., 1999) and rbcL (Groot et al., 2011) to be quite well conserved so the method is not appropriate for all barcode sequences. Similarly, local pairwise alignments are suitable for distantly related sequences since they find local regions of similarity.

Once MSA has been obtained, the alignment can be manually edited to increase alignment quality using BioEdit, Jalview or any other alignment editor followed by the construction of phylogenetic trees. Trees can be constructed by either distance based or character based methods. Distance based methods include Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sneath and Sokal, 1973) and neighbor-joining (NJ) (Saitou and Nei, 1987), whereas the character based methods are parsimony (Eck and Dayhoff, 1966) and maximum likelihood (Felsenstein, 1981).

Distance based methods (Chase et al., 2005, Taylor and Harris, 2012) rely on the DNA sequence variation between and within the species. TaxI (Steinke et al., 2005) is one such tool for DNA barcoding based on distance methods. Though the method is suitable for large datasets and for further analysis and also permits lineages with different branch lengths, it can be misleading in the accurate assignment of a query sequence to a particular taxa because it assumes that intra-specific variation is less than inter-specific variation, which may not always be the case, e.g. amphibians (Meyer and Paulay, 2005).

Character based DNA barcoding, is based on the concept that members of a given species share sequence attributes that are absent from members of other sister species (Kress et al., 2005; Hollingsworth et al., 2011). This approach characterizes species on the basis of the presence and absence of unique diagnostic characters, the four standard nucleotides (A, T, C, G). The success of character based DNA barcoding approach depends on the efficient identification of the diagnostic characters in the hierarchically organized DNA sequences (like in a phylogenetic tree). Moreover, to check the reliability of the identified character states, the ability of the character states to identify the query sequences needs to be tested. Examples of such softwares are the CAOS (Characteristic Attributes Organization System), Package (Sarkar et al., 2008) and BRONX (Little, 2011).

There are also several new computational methods in DNA barcoding: Compensatory base changes (CBCs), Operational Taxonomic Units (OTUs), DNA metabarcoding, psbA-trnH Specific barcoding, Representing DNA barcodes, Neutral networks, Machine learning, Data mining, Composition vector (CV) method. For more details about these methods, see the review of Bhargava and Sharma, 2013.

Apart from the kind of marker used, DNA barcoding does not require an extensive knowledge of the genome of each organism, being based on the use of one or few universal DNA sequences (Hollingsworth et al., 2011). On the contrary molecular based approaches (i.e. hybridization-based markers and PCR-based markers) are highly species-specific and their application is often limited to a single taxon, or to closely related taxa.

The limits of adopting universal barcode markers are evident at the cultivar level, where genetic variability is limited, and there are complications due to breeding events. To overcome these limits, Kane and Cronk (2008) proposed the ultra-barcoding methodology, which is based on the sequence of the whole plastidial genome, together with large portions of the nuclear genome. This combination provides enough information to evidence genetic diversity below the level of species, distinguishing hybrids from pure lines, hence it is far more sensitive than traditional DNA barcoding (Nock et al., 2011; Parks et al., 2009; Steele and Pires, 2011). Kane and Cronk (2008) evaluated the effectiveness of ultra-barcoding on cocoa (*Theobroma cacao* L.), and found several plastidial and nuclear SNPs, which were useful to identify different cultivars. This technique is promising, but it is difficult to apply on a large scale due to its high costs and its excessive species-specificity. Nowadays, the reduced genetic diversity at cultivar level often requires the analysis of large portions of the genome, which currently have a too high cost/effectiveness ratio to be widely used. Furthermore, this approach is contrary to the basic DNA barcoding methodology, which requires the analysis of short and universal DNA regions only.

2.2 The uses of DNA barcoding

From its inception, the primary use of DNA barcodes has been for species identification (Herbert et al., 2003). As a research tool for taxonomist, barcoding assist in identification by expanding the ability to diagnose species by including all life history stages of an organism (i.e. seeds, seedlings, eggs, larvae, mature individuals both fertile and sterile), unisexual species, damaged specimens, gut contents, scat samples. In addition, systematists have the potential to quantify the consistency of their species definition with a universal measure of genetic variability based on the barcode data.

DNA barcodes are now also being used to address fundamental ecological and evolutionary questions, such as how species in plant communities are assembled (Kress et al., 2009; Kress et al., 2010) and the degree of specialization in tropical versus temperate zone herbivores (Rivera et al., 2009). To date DNA barcoding is a powerful tool in many fields below some practical applications to ensure high quality standards for food industry and market.

DNA barcoding was proven to be particularly effective in the traceability of seafood (Becker. et al., 2011). To date, more than 70,000 barcode sequences from 8300 species (26% of the total) have been stored in the framework of an international collaborative research: the Fish Barcode of Life Initiative (FISH-BOL—www.fishbol.org). FISH-BOL represents one of the most comprehensive resources for the analysis of fish and seafood products (Ward et al., 2009).

For traceability of milk and dairy products (i.e. foodstuffs made from mammalian milk) the plastidial rbcL, the most universal marker for plant DNA barcoding, is able to detect traces of free-derived plant DNA fragments in raw milk and in its fractions (Ponzoni et al., 2009).

Bruni et al. (2010) evaluated the effectiveness of DNA barcoding in separating toxic from edible species, evidencing a clear molecular distinction between cultivated species of the genera *Solanum* (*Solanum* *tuberosum* L., *Solanum lycopersicum* L. group) and *Prunus* (*Prunus armeniaca* L., *Prunus avium* L., *Prunus cerasus* L., *Prunus domestica* L.) and their toxic congenerics. This study suggested that DNA barcoding could be used to distinguish edible species from their non-edible or toxic congenerics (Jaakola et al., 2010).

DNA barcode markers were also efficiently used to identify commercial tea (Stoeckle et al., 2011), fruit species in yogurt (Knight et al., 2007), and fruit residues in juices, purees, chocolates, cookies, etc. (Sakai et al., 2010).

DNA barcoding showed a high effectiveness in the evaluation of the presence of allergenic species, both in fresh and in processed food. Nuts are considered one of the main sources of allergens (Hubalkova and Rencova, 2011), and their presence in food (also in traces) is detectable by molecular analysis based on different markers, including DNA barcode regions (Yano et al., 2007). The identification of allergenic material is one of the more interesting applications of DNA barcoding. It can be used to satisfy the requirements of FAO and European Commission, which list allergenic species that must be declared on food labels (Directive 2003/89/EC.1).

Due to its universality, DNA barcoding can be used in different contexts and by different operators. International agencies or institutions, which are responsible for quality control of raw materials or food commodities, can cooperate by exchanging their data, hence creating population reference databases, the lack of which is the only real limit of the method. In fact, while some groups of organisms (e.g. fish) are well represented and provide a reliable source of reference DNA barcoding, others have been poorly investigated. For this reason, in the near future DNA barcoding is likely to become a routine test in many fields, such as food quality control and traceability.

2.3 Our barcoding approach

The objective of our research is the identification of inbred and Lombard varieties of *Zea mays*, the latter characterized by individuals genetically similar as result of their common geographical origin. For this purpose, we chose as a marker the α-zein genes. As describe above these genes show an extreme heterogeneity at both DNA and amino acid level, which are desirable characteristics for a marker able to be discriminated among closely related cultivars. Moreover in our laboratory there is a consolidate experience with the α-zein system, thus, facilitating their investigation. Finally, these genes are universally present in the genome of all maize lines and cultivars and, therefore, universal primers for barcoding can be generated. In other words α -zein genes satisfy all three criteria established by the CBOL.

To this purpose, we have identified two alternative approaches for the identification and discrimination of closely related individuals of *Zea mays* among inbred lines and cultivars coming from the same geographical area (Piccinini et al., 2014).

1. DNA barcoding based on molecular marker Restriction Fragment Length Polymorphism (RFLP). RFLP profiles, derived from a combination of restriction enzymes and α-zeins as probes, were able to produce consistent restriction fragment patterns which can be converted into a binary code (0-1) and then into a barcode using the online software.
2. Protein Barcoding. We produced a 2D profile of storage proteins (in particular α-zeins) of each inbred and maize varieties. Gels were scanned in an Epson Expression 1680 Pro Scanner and analysed with ImageMaster 2-D Platinum Software v6.0 (GE Healthcare Life Sciences, USA). Automatic matching was complemented by manual matching. Molecular weights of the spots were estimated using the migration range standard as reported in Viotti et al. (1982), while pI was determined as described in Righetti et al., 1977. The presence or absence of a spot generates the corresponding barcode.

With both these approaches, each genotype can be univocally identified. Even if these methods may seem far from those above mentioned, RFLP-barcoding has in common with the concept of DNA barcoding, proposed by Herbert, the use of a short sequence of DNA (zeins sequence) which satisfy the tree criteria of universality, sequence quality and discrimination.

Aims of the project

This PhD project has two aims:

1. The first was to establish a data bank of zein expressed genes coding for the α-22kDa polypeptides in order to develop a suitable strategy for producing artificial zein genes encoding for polypeptides with a higher content in lysine and methionine. These synthetic genes, like their wild-type counterpart, have to be sorted and correctly accumulated in the lumen of RER, and must be assembly to form proper protein bodies;

2. The second was to exploit the data retrieved from the extensive activity of zein sequences cloning. These data are used as an *intra*-species recognition tool, individual barcode, based on zein-Proteomic-Genomic-Profiles, zPGP, to be used for inbreds and Lombard varieties discrimination.

Main results

Foreword

Three major results were achieved in the present work with the aims to establish approaches suitable to identify accessions of maize genotypes and to validate intracellular behaviour of modified zein coding sequences into polypeptides with balanced amino acid composition. For the latter approach zein and intra-species sequences were considered in view of transformation approaches based on cisgenesis.

1. Indel and single nucleotide variations of zeins generate unique 2D-zein patterns and molecular markers suitable for Zea mays genotyping

Lombard varieties and maize inbreds barcoding were obtained by 2D gel and Southern blot analyses. For each genotype the 2D and Southern blot patterns were converted into a binary code, and then into a barcode. In both approaches, each genotype was univocally identified making zeins a valuable tool for rapid identification of several Lombard varieties and different maize lines at individual level. For more details see the published paper in the second part of this thesis.

2. Glutelin genes are suitable tools for maize (Zea mays) genotyping

To use glutelin sequences as molecular markers for genotyping, we designed glutelin-specific primers for each glutelin prototype (G1L, G1S, G2 and G3). Primers for the G1L, G1S, G2 and G3 were designed at the 5' and at the 3' ends of their coding sequences (Table1). In Figure 1 the amplification results of each primer combination are shown together with the results of their hybridization specificity as determined by DNA gel blot analysis. As G1L and G1S sequences have a high identity at the 3' end, we designed two different reverse primers, specific for each sequence, to obtain fragments close to the 5' end.

Primers	**FORWARD**	**REVERSE**
G1L	CGATCGACACCATGAGGGTG	ATAGTTTCTTCAGTGGGGGA
G1S	ACTCGACACCATGAAGGTGC	TACATAGTTTCCTCAGTAGTAGAC
G2	AACAGAACAGCATGAAGATGGTCATC	CCCAAATATCATGAATCAGTAGTAGG
G3	CCGCCATGGCAGCCAAGATG	TCTATCTAGAATGCAGCACCAAC

Tab. 1: Primers used in the amplification of glutelin genes. The start and the reverse complementary stop codons are underlined

The entire PCR products of G2 and G3, and the fragments of G1L and G1S were used as probes to characterize the extent of DNA polymorphisms at the glutelin loci among subset of Lombard maize varieties and inbred lines. DNAs were extracted from sporophytic tissue and digested with the HindIII and DraI restriction enzymes. These enzymes were chosen for the analysis as they do not digest within the glutelin coding region and are insensitive to DNA methylation. Each probe-restriction enzyme combination produced DNA blot patterns that were specific for each glutelin (Figure 1 B). DNA blots probed with G2 or G3 or even with the fragments of G1L or G1S, evidenced two to three major bands indicating the occurrence of at least two genes in most of the genotypes.

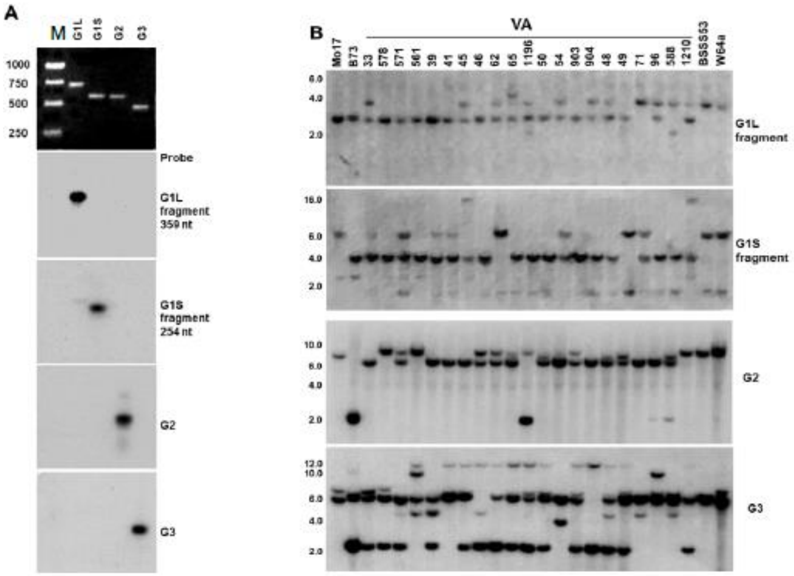


Fig. 1 Probe specificity and genomic DNA gel blot analyses of glutelin genes.

A) PCR products of the 4 glutelin prototypes were run in parallel and then transferred to nylon membranes. Each filter was hybridised to glutelin probes as indicated on the right. The upper panel reports size markers in bps as indicated on the left and the fractionation of the amplified fragments (G1L, G1S, G2 and G3 as indicated on the top) obtained with primers reported in tab. 1.B) Genomic DNAs were digested with DraI in the two upper panels and with HindIII in the two lower panels. Filters were hybridised to probes as indicated on the right. Molecular mass markers in kb are on the left.

Indeed, the analysis of the entire set of Lombard varieties with these probes and restriction enzyme combinations permitted to generate on the basis of presence/absence of the various bands a binary code, which was transformed into a barcode that identified univocally the twenty-five genotypes. The heatmap depicts the uniqueness and creates a relationship among the various genotypes (Figure 2).

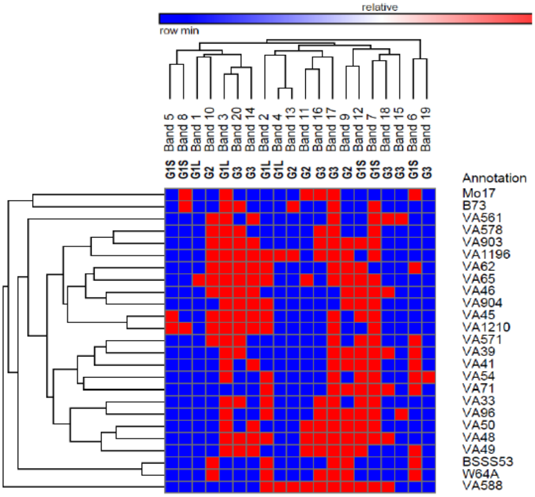


Fig. 2: Hierarchical clustering of the RFLP bands presence/absence, by DNA blot analysis.

3. Modified zein polypeptides with higher content in methionine and lysine are accumulated into the RER lumen of tobacco protoplasts

3.1 Architecture of the synthetic genes

The N-terminal sequence of G3 is joined to the C-terminal of H3, or H4, by a short stretch of six amino acids (Glycine-Alanine-Glycine-Serine-Glycine-Glycine=GAGSGG) which ensures conformational freedom of two domains. In Figure 3 the amino acid and nucleic acid sequences of G3H3 and G3H4 are reported.

(A) G3H3

MAAKMLALFALLALCASATSA

THIPGHLPPVMPLGTMNPCMQYCMMQQGLASLMACPSLMLQQLLALPLQTMPVMMPQMMTPNMMS
PLMMPSMMSPMVLPSMMSQIMMPQCHCDAVSQIMLQQQLPGAGSGGGTVALREIRKYQKSTELLI
RKLPFQRLVREIAQDFKTDLRFQSSAVAALQEAAEAYLVGLFEDTNLCAIHAKRVTIMPKDIQLA
RRIRGERA* 203 aa 6xK 25xM

ATGGCAGCCAAGATGCTTGCATTGTTCGCTCTCCTAGCTCTTTGTGCAAGCGCCACTAGTGCGAC
CCATATTCCAGGGCACTTGCCACCAGTCATGCCATTGGGTACCATGAACCCATGCATGCAGTACT
GCATGATGCAACAGGGGCTTGCCAGCTTGATGGCGTGTCCGTCCCTGATGCTGCAGCAACTGTTG
GCCTTACCGCTTCAGACGATGCCAGTGATGATGCCACAGATGATGACGCCTAACATGATGTCACC
ATTGATGATGCCGAGCATGATGTCACCAATGGTCTTGCCGAGCATGATGTCGCAAATAATGATGC
CACAATGTCACTGCGACGCCGTCTCGCAGATTATGCTGCAACAGCAGTTACCAGGTGCTGGATCA
GGTGGTGGCACCGTCGCGCTCCGGGAGATTCGCAAGTACCAGAAGAGCACGGAGCTGCTCATCCG
CAAGCTGCCCTTCCAGCGCCTCGTCCGTGAGATCGCGCAGGATTTCAAGACCGACCTCCGCTTCC
AGTCCTCCGCTGTCGCCGCGCTGCAGGAGGCCGCCGAGGCCTACCTCGTGGGGCTCTTCGAGGAC
ACCAACCTCTGCGCCATCCACGCCAAGCGCGTCACCATCATGCCCAAGGACATCCAGCTCGCGCG
CCGCATCAGGGGCGAGAGGGCT**TGA** 675bp

(B) G3H4

MAAKMLALFALLALCASATSA

THIPGHLPPVMPLGTMNPCMQYCMMQQGLASLMACPSLMLQQLLALPLQTMPVMMPQMMTPNMMS
PLMMPSMMSPMVLPSMMSQIMMPQCHCDAVSQIMLQQQLPGAGSGGDNIQGITKPAIRRLARRGG
VKRISGLIYEETRGVLKIFLENVIRDAVTYTEHARRKTVTAMDVVYALKRQGRTLYGFGG*
190 aa 5xK 25xM

ATGGCAGCCAAGATGCTTGCATTGTTCGCTCTCCTAGCTCTTTGTGCAAGCGCCACTAGTGCGAC
CCATATTCCAGGGCACTTGCCACCAGTCATGCCATTGGGTACCATGAACCCATGCATGCAGTACT
GCATGATGCAACAGGGGCTTGCCAGCTTGATGGCGTGTCCGTCCCTGATGCTGCAGCAACTGTTG
GCCTTACCGCTTCAGACGATGCCAGTGATGATGCCACAGATGATGACGCCTAACATGATGTCACC
ATTGATGATGCCGAGCATGATGTCACCAATGGTCTTGCCGAGCATGATGTCGCAAATAATGATGC
CACAATGTCACTGCGACGCCGTCTCGCAGATTATGCTGCAACAGCAGTTACCAGGTGCTGGATCA
GGTGGTGACAACATCCAGGGCATCACCAAGCCGGCGATCCGGAGGCTGGCTAGGAGGGGTGGCGT
GAAGCGCATCTCGGGGCTCATCTACGAGGAGACCCGCGGCGTGCTCAAGATCTTTCTCGAGAACG
TCATCCGCGACGCCGTCACCTACACCGAGCACGCGCGCCGCAAGACCGTGACCGCCATGGACGTC
GTCTACGCGCTCAAGCGCCAGGGCCGCACCCTCTACGGCTTCGGAGGC**TAG** 636bp

Fig. 3: Amino acid and nucleic sequence of G3H3 (A) e G3H4 (B). In bold, the 21 aa$_s$ sequence of signal peptide; in yellow the 105 aa$_s$ sequence of G3 NH2-terminal;

in green the amino acid linker between the two domain and in blue the H3 92 aa$_s$ and the H4 79 aa$_s$ C-terminal. (*) corresponds to the stop codon indicated in bold in the nucleic acid sequence.

As mentioned in the introduction of this thesis, for ZRK we adopted a different strategy, namely the modification of the coding capacity of an alpha-zein gene was obtained by changing arginine codons into lysine codons. Figure 4 highlights the lysine residues which replace arginine (R) present in the wild-type gene.

```
MAAKIFAILALLALSASVATA
TIIPQCSQQYLSPVTAAKFEYPTIQSYKLQEAIAASILKSLALTVQQPYALLQQPSLVN
LYLQKIVAQQLQQQLLPTINQVVAANLDAYLQQQQFLPFNQLAGVNPAAYLQAQQLLPF
NQLVKSPAAFLLQQQLLPFKLQVVANIAAFLQQQQLLPFYPQVVGNINAFLQQQQLLPF
YPQDVANNVAFLQQQQLLPFSQLALTNPTTLLQQPTIGGAIF* (219aa 6xK)
```

```
ATGGCAGCCAAGATTTTTGCCATCCTTGCCCTCCTTGCTCTTTCAGCAAGCGTTGCTACCGCGAC
TATTATTCCACAATGCTCACAACAATACCTCTCTCCGGTGACAGCCGCGAAATTTGAATACCCAA
CTATACAATCCTACAAGCTACAAGAGGCCATCGCAGCAAGCATCTTAAAGTCGTTAGCATTGACT
GTCCAACAACCATATGCCCTATTGCAACAACCATCCTTAGTGAATCTATATCTCCAAAAGATCGT
AGCACAACAACTACAACAACAATTGCTTCCAACAATCAATCAAGTAGTTGCAGCGAACCTTGATG
CTTACCTCCAGCAACAACAATTTCTTCCATTCAATCAACTAGCTGGGGTGAACCCTGCTGCTTAC
TTGCAGGCACAACAGCTACTACCATTCAACCAACTTGTCAAGAGCCCTGCTGCCTTCTTACTGCA
GCAACAGTTGTTGCCATTCAAACTACAAGTTGTGGCAAACATTGCTGCTTTCTTGCAACAACAAC
AATTGCTGCCATTTTACCCACAGGTTGTGGGAAACATTAACGCCTTCTTGCAACAGCAACAGTTG
CTGCCATTCTACCCACAGGATGTGGCAAACAATGTCGCCTTCTTACAACAACAACAATTGCTGCC
ATTTAGCCAACTTGCTTTGACGAATCCTACCACCTTATTGCAGCAGCCCACCATTGGTGGTGCCA
TCTTTTAG 723 bp
```

Fig. 4: Amino acid and nucleic sequences of ZRK. In bold, the 21 aa$_s$ sequence of signal peptide. (*) corresponds to the stop codon indicated in bold in the nucleic acid sequence. Highlighted in blue the lysines which replace the arginine residues present in the wild-type gene. The triplets coding for lysines are underlined.

In the table below we compare the properties of a typical mature α- zein wild type (wt) against the mature synthetic proteins (Table 2).

Gene	N° amino acids of mature polypeptide	Molecular weight (kDa)	Isoelectric point (pI)	Content of M or C (mol %)	Content of K (mol %)
α-zein wt	219	24.5	7.07	0.5	0.0
G3H3	203	22.6	9.03	15.2	3.9
G3H4	190	21.0	9.62	15.8	3.2
ZRK	219	24.4	8.68	0.5	2.7

Tab. 2: Proprieties of synthetic and wild-type zeins.

3.2 In vitro transcription and translation of new genes and processing of their polypeptides

The eukaryotic Rabbit Reticulocyte Lysate added with the canine membranes were used as an in *vitro* system to verify the transcription-translation of the synthetic genes and the proper processing of the synthesized polypeptides. In particular, the system ensures the cleavage of the signal peptide and post-translational modifications. The three constructs were obtained by cloning into the pBS vector (Figure 5) the inserts of interest in the XbaI and HindIII sites downstream the T7 promoter. The 5 'UTR sequence of G3 was inserted upstream of the coding sequence of G3H3 and G3H4, whereas the 5 'UTR of an α-zein gene was inserted upstream of the ZRK coding sequence. All constructs end with a poly (A) tail. The addition of this element ensures, in *vitro*, the stabilization of the transcripts and its translational efficiency.

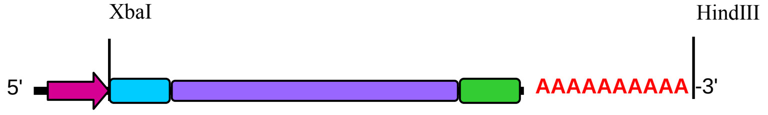


Fig. 5: Schematic drawing and cloning sites of pBS vector.

Starting from the 5' end the coloured parts represents: the pink arrow the T7 promoter; the blue rectangle the 5' UTR sequence of G3 for G3H3 and G3H4 or the 5' UTR sequence of an α-zein for ZRK; the purple rectangle the gene of interest; the green rectangle the 3' UTR sequence of G3 for G3H3 and G3H4, or 3' UTR sequence of an α-zein for ZRK. At the end the polyA tail.

The HindIII-linearized recombinant pBS plasmids were transcribed with the T7 phage RNA polymerase in the presence or absence of7-methylguanosine triphosphate to obtain cap-plus or cap-minus transcripts, the former known to have higher efficiency than the latter in translational experiments. Both types of transcripts were used in *in vitro* translational experiments with similar results in polypeptide synthesis and processing.

About 100 ng of mRNA were added to the Rabbit Reticulocyte system in the absence or in the presence of canine membranes. Polypeptide synthesis was monitored by addition of [^{35}S] methionine and cysteine. Figure 6 shows the fractionation by SDS-PAGE of total proteins produced by translation of the cap-minus transcripts of G3H4, G3H3 and ZRK. Referring to Figures 3 and 4, and to Table 2, the proteins produced in this *in vitro* system have molecular masses that reflect the expected ones derived from the cleavage or non-cleavage of the signal peptide.

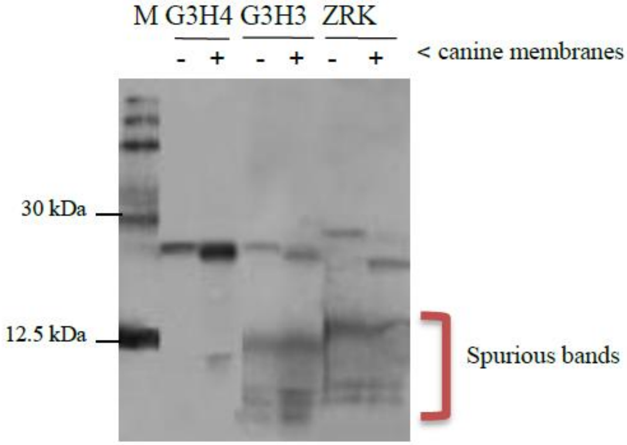


Fig. 6: Analysis by SDS-PAGE of gene products of clones of interest obtained by *in vitro* translation with the Rabbit Reticulocyte Lysate system in absence (-) or presence (+) of canine membranes. Gel was dried and exposed to X-ray film for autoradiography. M radioactive protein markers, molecular weights are indicated on the left.

3.3 Expression of synthetic and wild type zeins in protoplasts of *Nicotiana tabacum*

The above results indicate that the coding region of the 3 synthetic genes contains the correct information to produce mature proteins with the proper cleavage of the signal peptide. The next steps were: i) to verify in an *in vivo* plant derived system, leaf tobacco protoplasts, the correct processing and ii) to control the intracellular localization and accumulation of the various proteins. The synthetic genes and the wild-type genes (G1L, G1S, G2) were cloned in the BamHI-SalI sites of the pDHA vector (Figure 7) downstream the 35S constitutive promoter of the Cauliflower Mosaic Virus (CaMV35S) and the leader sequence of the alfalfa mosaic virus (AMV). Downstream the 3' end of the genes the 35S transcriptional terminator is also present.

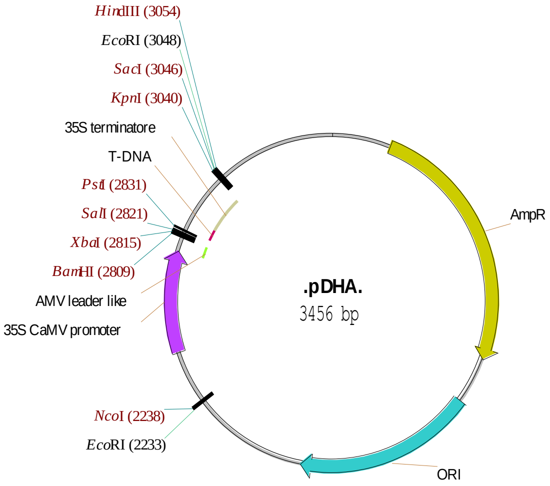


Fig. 7: Map of pDHA vector.

After protoplast transformation the proteins obtained were analysed on SDS-PAGE gels and characterized by using polyclonal antibodies specific for the protein products. As known, zeins and glutelins are soluble in alcoholic solvents, whereas, from our preliminary investigation, G3 recombinant polypeptides because of the histone moieties became insoluble. Accordingly, Figure 8A and 8B show the results of Western Blot analysis of proteins soluble and insoluble in 70% ethanol, respectively.

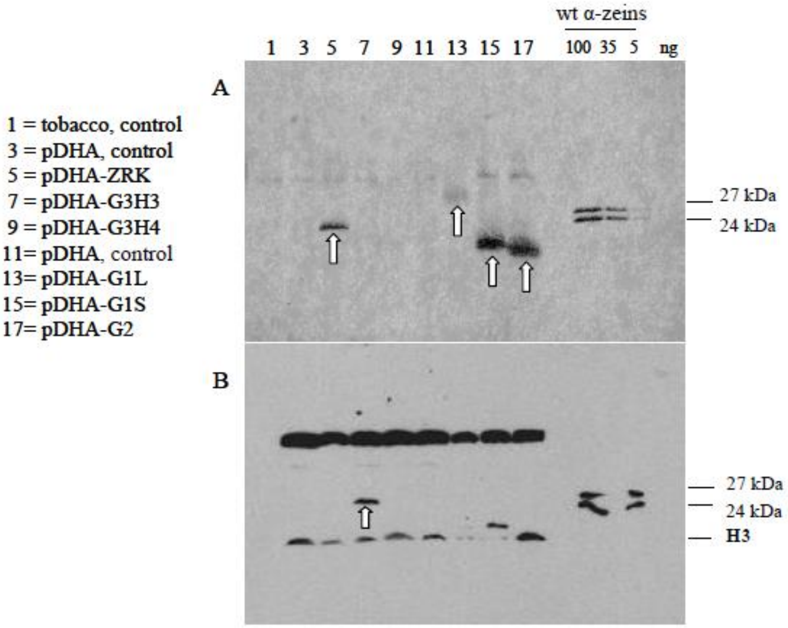


Fig. 8: Expression in *in vivo* plant derived system.

SDS-PAGE of protein extracts obtained from protoplasts of *N. tabacum* transformed with the genes of interest. The last three lanes on the right were loaded with different concentrations of α-zein. On the right and on the top numbers indicate the transformation events and apply for both panels. Proteins were blotted to membrane filters. The white arrows indicate the protein product of interest displayed with specific antibodies. A) The membrane filter was incubated with a mixture of antibodies that recognize α-zein and glutelins. B) The filter was treated first with α-zeins antibody, then stripped, and secondly incubated with H3 antibody. To the right the molecular weights of wild type zeins are reported. H3 indicate the position of histone-3 derived from tobacco protoplasts.

As the synthetic constructs and wild-type genes were correctly expressed *in vivo*, apart from some difficulty in detecting the G3H4 protein (data not shown), the following experiments were aimed to investigate the intracellular localization of the protein products. In this regard we carried out a second transformation of tobacco protoplasts.

Transformed protoplasts were centrifuged on a sucrose gradient to fractionate cellular organelles (vacuoles, ER, nuclei).

Figure 9 reports the distribution in the different cellular compartments of the synthesised proteins. Using an anti-Bip serum (a protein specific of the lumen of the endoplasmic reticulum and involved in protein folding) it was possible to identify the fractions in which the endoplasmic reticulum bands and to obtain biochemical evidence of intracellular localization of the proteins of interest. All the proteins are recovered in the ER fractions.

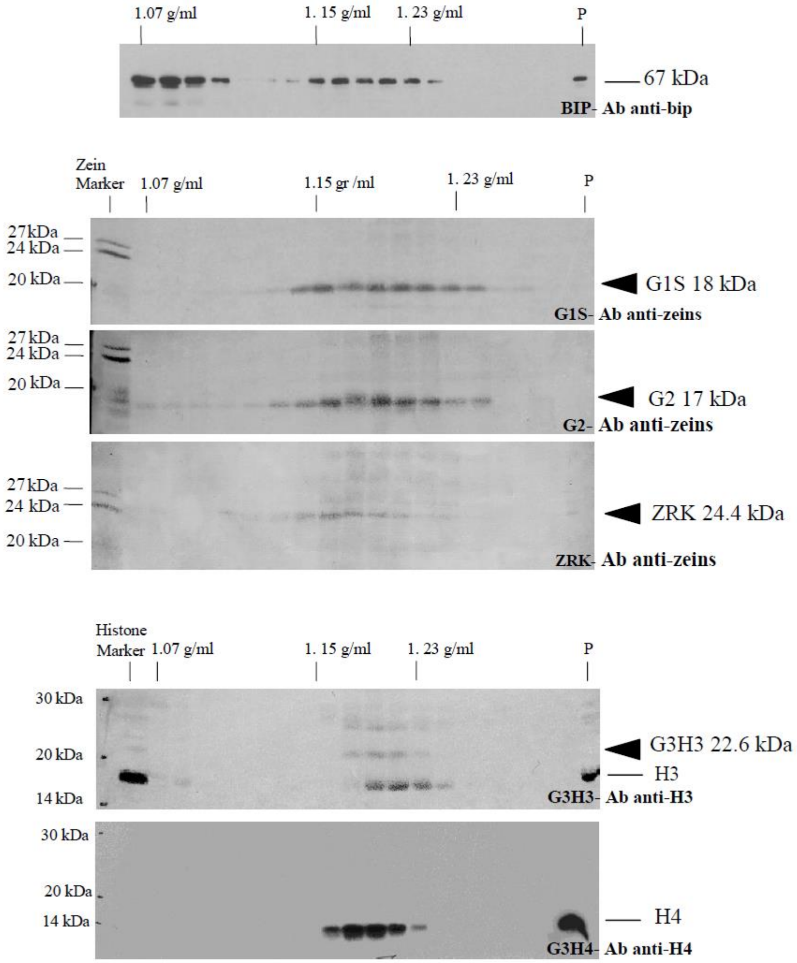


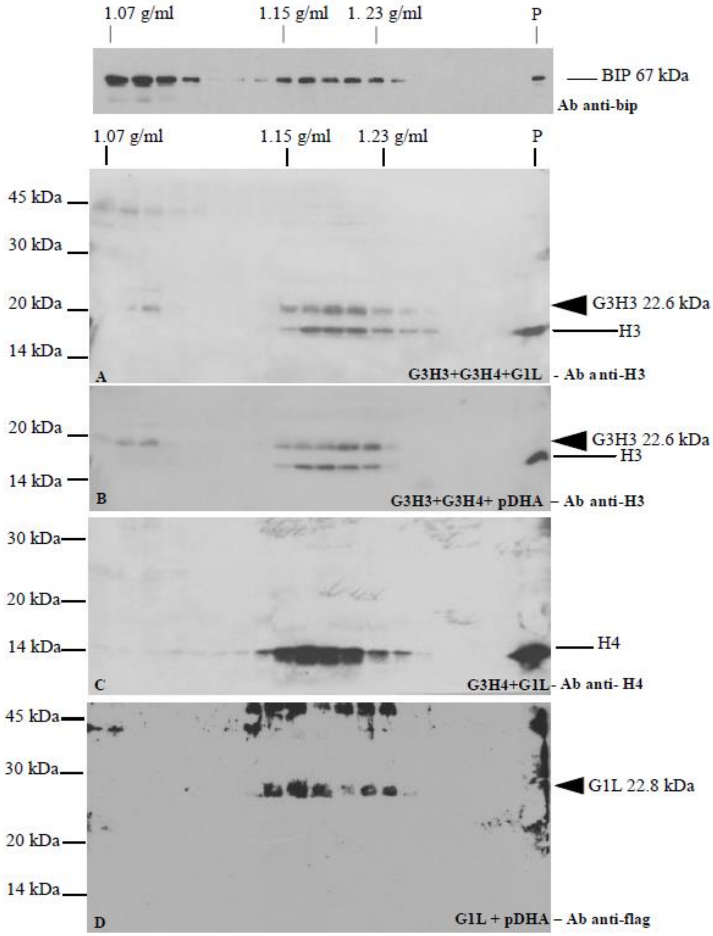
Fig. 9: Subcellular fractionation of transformed protoplasts.

Protoplasts were homogenised in EDTA buffer and subjected to centrifugation on a 16-55 % (w/w) isopycnic sucrose gradient. Gradient fractions were collected and resolved by SDS-PAGE followed by protein blot to membrane filters. Filters were decorated with anti-BiP, anti-zeins or anti-histones immunsera. The number at the top indicates fraction density in g/mL, ER bands at a density of 1.15-1.23 g/ml. Numbers on the left indicate protein molecular markers. P is the pellet fraction formed during centrifugation. The dark arrowheads on the right indicate the molecular mass of the proteins of interest.

The successive step was the co-expression, in tobacco protoplast, of the synthetic genes in different combination with or without the G1L wild-type gene known to help the protein body formation. In total we carried out six co-transformations using 15 ng of DNA for each construct:

1. G1L + ZRK;
2. G1L + pDHA;
3. G1L + G3H3 + G3H4;
4. G3H3 + G3H4 + pDHA;
5. G1L + G3H3 + G3H4 + ZRK;
6. G1L + G3H4.

In this case the G1L construct has a C-terminal amino acid flag (Asp-Tyr-Lys-Asp-Asp-Asp-Asp-Lys) that facilitates its identification through a monoclonal antibody specific for this sequence. As for the experiments previously conducted with the antibody anti-Bip, it was possible to identify the gradient fractions in which the endoplasmic reticulum is present. For the identification of other proteins we used a series of rabbit polyclonal antisera as indicated in Figure 9. The products of synthesis and accumulation co-localize with Bip (Figure 10) in a density range between 1.15 and 1.23 g/ml typical of the ER.


1.07 g/ml
1.15 g/ml
1. 23 g/ml
P
BIP 67 kDa
Ab anti-bip
1.07 g/ml
1.15 g/ml
1.23 g/ml
P
45 kDa
30 kDa
20 kDa
14 kDa
G3H3 22.6 kDa
H3
A
G3H3+G3H4+G1L - Ab anti-H3
20 kDa
14 kDa
G3H3 22.6 kDa
H3
B
G3H3+G3H4+ pDHA – Ab anti-H3
30 kDa
20 kDa
14 kDa
H4
C
G3H4+G1L- Ab anti- H4
45 kDa
30 kDa
G1L 22.8 kDa
20 kDa
14 kDa
D
G1L + pDHA – Ab anti-flag

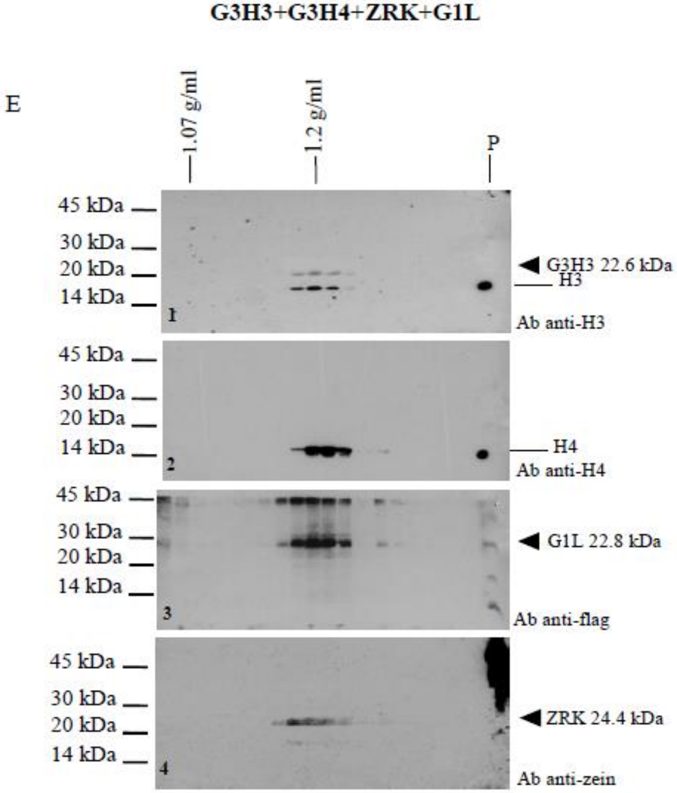


Fig. 10: Subcellular fractionation of transformed tobacco protoplasts.

For each transformation tobacco protoplasts were homogenised in EDTA buffer and subjected to centrifugation on a 16–55 % (w/w) isopycnic sucrose gradient. Gradient fractions were collected and resolved by SDS-PAGE. Proteins were blotted to membrane filters and immunodecorated with antibodies as indicated on the bottom right of each filter. The numbers at the top indicate fraction density in g/ml. P is the pellet formed during centrifugation, which contains nuclei. Numbers at the left indicate molecular mass markers in kilo Daltons. Bip has a real molecular weight of 78 kDa, but its apparent molecular weight is 67 kDa. The panels from A to D report on the bottom right the constructs utilised in the transformation. On the right of each panel, an arrowhead points to the polypeptide expected from the transformation process. For A and B panels, H3 indicates the H3-histone recovered in the nuclei and ER fractions. For panel C, H4 points to H4-histone recovered in the nuclei and ER fractions. Panels E report the detection of the proteins of interest from the transformation carried out with the four constructs, G1L + G3H3 + G3H4 + ZRK. After protein blotting, the filter was

treated from 1 to 4 with antisera, as indicated on the right, and with stripping after each immunodecoration.

Collectively, the results reported in Figures 8, 9 and 10 indicate that the three genes of interest produce polypeptides that: i) are properly processed for the cleavage of the signal peptide; ii) processing occurs on ER membranes and iii) after processing the mature polypeptides are accumulated in the ER lumen. However, a particular feature occurs for the G3H3 and G3H4 constructs. For the former, the entire recombinant protein is recovered into the ER lumen together with a fragment that has a relative mobility identical to that of H3-histone (Figure 9-G3H3, and Figures 10A and 10E1). For G3H4, only a fragment that migrates as the H4-histone is revealed (Figure 9-G3H4, and Figures 10C and 10E2). The recovery of the two fragments indicates that some specific cleavage by an endopeptidase occurs, however, only partially for G3H3.

In Figures 9 and 10, the signals of the various polypeptides, apart from ZRK, were obtained with short exposure time. Detection of ZRK requires several minutes of exposure indicating that the translation of this mRNA, due to its particular codon usage, may be impaired by different tRNA pools present in maize endosperm and tobacco protoplast (Viotti et al., 1978). The results of the co-transformation of the three constructs, in presence or absence of G1L, do not evidence a higher expression of the three recombinant proteins, as reported in other published reports. In any case, with regard to wild-type zeins or glutelins the recombinant polypeptides with a balanced amino acid composition are accumulated into the ER lumen.

Conclusions and future prospects

The approaches developed and utilised in this research were instrumental to the genotyping of each maize genotype allowing to identify at individual levels both varieties and inbred lines.

The extensive analyses of both DNA and proteins of zein and glutelin sequences produce dendograms with different relationships among the maize genotypes since they are based on inter genic variations (RFLP) and gene expression (2D). However, for the eight accessions of B73 line we evidenced an almost constancy of pattern in their RFLP, whereas SNPs were revealed in the coding sequences, which in any case do not determine differences in their 2D pattern. The SNPs revealed in the B73 accessions and the other SNPs recovered by nucleotide comparison of *in house* and in published datasets of zein genomic and cDNA sequences, allowed to create a specific set of primers to develop a PCR multiplex approach, now in progress for a simpler and faster DNA analysis.

Similarly, the expression analysis of both wild-type and synthetic genes was instrumental in defining that all the signal peptides of the zeins and glutelins used, properly targeted to ER the various constructs. Moreover, even though with some difference in the amount recovered and even in their stability, the proteins of interest, with a balanced amino acid composition, were shown to be retained in the ER lumen. The above results support the adopted strategy of modifications, and may indicate further steps of manipulation to increase both stability and accumulation. One such approach could be based on the above mentioned evidence of proper zein and glutelin signal peptide targeting to ER. The constructs of the fused proteins would consider the C-terminal H2A and H2B histones to stabilize a structure similar to that occurring in the nucleosome. Retention in the ER should be provided by the zein or glutelin N-terminal part as already demonstrated for other recombinant proteins.

References

Altenbach SB, Kuo CC, Staraci LC, Pearson KW, Wainwright C, Georgescu A, Townsand J. 1992. Accumulation of brazil nut albumin in seeds of transgenic canola results in enhanced levels of seed protein methionine. *Plant Molec. Biol.*18,. 235-245.

Argos P, Pedersen K, Marks MD, Larkins BA 1982. A structural model for maize zein proteins. *Journal Biological Chemistry.*257, 9984-9990.

Arruda P, Bright SWJ, Kueh JSH, Lea PJ, Rognes SE. 1984. Regulation of aspartate kinase isoenzymes in barley mutants resistant to lysine plus threonine—construction and analysis of combinations of the *LT1A*, *LT1B*, and *LT2* mutant-genes. *Plant Physiol.* 76, 442-446.

Azevedo RA, Arruda P, Turner WL, Lea PJ. 1997. The biosynthesis and metabolism of the aspartate derived amino acids in higher plants. *Phytochemistry*. 46, 395-419.

Azevedo RA, Lea PJ . 2001. Lysine metabolism in higher plants. *Amino Acids*. 20, 261-279.

Azevedo RA. 2002. Analysis of the aspartic acid metabolic pathway using mutant genes. *Amino Acids*. 22, 217-230.

Azevedo RA, Damerval C, Landry J, Lea PJ, Bellato CM, Meinhardt LW, Le Guilloux M, Delhaye S, Toro AA, Gaziola SA, Berdejo BD. 2003. Regulation of maize lysine metabolism and endosperm protein synthesis by opaque and floury mutations. *Eur. J. Biochem.* 270 (24), 4898-4908.

Azevedo RA, Lancien M, Lea PJ. 2006. The aspartic acid metabolic pathway, an exciting and essential pathway in plants. *Amino Acids*. 30, 143-162.

Bicar EH, et al. 2008. Transgenic maize endosperm containing a milk protein has improved amino acid balance. *Transgenic Research.* 17, 59-71.

Boulter D, Croy RRD. 1997. The structure and biosynthesis of legume seed storage proteins: A biological solution to the storage of nitrogen in seeds. *Advances in Botanical Researches.* 27, 1-84.

Bright SWJ, Kueh JSH, Franklin J, Rognes SE, Miflin BJ. 1982. Two genes for threonine accumulation in barley seeds. *Nature*. 299, 278-279.

Burr B, Burr FA. 1976. Zein synthesis in maize endosperm by polyribosomes attached to protein bodies. *Proc. Nat. Acad. Sci USA.* 73, 515-519.

CBOL Plant Working Group. 2009. A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America.* 106, 12794-12797.

Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madrinan S, Peterson G, Seberg O, Jorgsensen T, Cameron KM, Carine M et al. 2007. A proposal for a standardised protocol to barcode all land plants. *Taxon.* 56, 295-299.

Coleman CE, Herman EM, Takasaki K, Larkins BA. 1996. The maize gamma-zein sequesters alpha-zein and stabilises its accumulations in protein bodies of transgenic tobacco endosperm. *Plant Cell*. 12, 2335-2345.

Coleman CE, Larkins BA. 1999. The prolamins of maize. *Seed Proteins.* 109-139.

Croy RRD, Hogue Ms, Gatehouse JA, Boulter D. 1984. The major proteins from pea (Pisum sativum L). Purification and some properties. *Biochemical Journal.* 218, 795-803.

Davison J. 2010. GM plants: Science, politics and EC regulations. *Plant Science*. 178 (2), 94-98.

Eggert LS, Rasner CA, Woodruff DS. 2002. The evolution and phylogeography of the Africa elephant inferred from mitochondrial DNA sequence and nuclear microsatellite markers. *Proceedings of the Royal society of London Series B-Biological Sciences*. 269, 1993-2006.

FAO. 2004. The state of Food and Agriculture 2003-2004. *Agriculture Biotechnology*.

Folmer O, Black M, Hoeh W, Lutz R, Vrijenhoek R. 1994. DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates. *Mol Mar Biol Biotechnol*. 5, 294-299.

Frizzi A, et al. 2008. Modifying lysine biosynthesis and catabolism in corn with a single bifunctional expression/silencing. *Plant Biotechnology Journal*. 6, 13-21.

Garratt R, Oliva G, Caracelli I, Leite A, Arruda P. 1993. Studies of the zein-like alpha-prolamins based on an analysis of amino acid sequences: implications for their evolution and three-dimensional structure. *Proteins*. 15, 88-99.

Geetha KB, Lending CR, Lopes MA, Wallace JC, Larkins BA. 1991. Opaque-2 modifiers increase c-zein synthesis and alter its spatial distribution in maize endosperm. *Plant Cell*. 3, 1207-1219.

Green CE, Phillips RL. 1974. Potential selection system for mutants with increased lysine, threonine, and methionine in cereal crops. *Crop Sci*. 14, 54-58.

Hamaker BR, Mohamed AA, Habben JE, Huang CP, Larkins BA. 1995. Efficient procedure for extracting maize and sorghum kernel proteins reveals higher prolamin contents than the conventional method. *Cereal Chem*. 72, 583-588.

Hebert PDN, Cywinska A, Ball SL, DeWaard JR. 2003a. Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Series B-Biological.* 270, 313-321.

Herbert PDN, Ratnasingham S, DeWaard JR. 2003b. Barcoding of animal life: cytochrome c oxidase subunit I divergences among closely related species. *Proceedings of the Royal Society of London Series B-Biological Sciences.* 270, S96-S99.

Hibberd KA, Green CE. 1982. Inheritance and expression of lysine plus threonine resistance selected in maize tissue culture. *Proc. Natl. Acad. Sci. USA.*79, 559-563.

Holding DR, Hunter BG, Chung T, Gibbon BC, Ford CF, Bharti AK, Messing J, Hamaker BR, Larkins BA. 2008. Genetic analysis of opaque2 modifier loci in quality protein maize. *Theor Appl Genet*. 117, 157-170.

Holding DR, Messing J. 2013. Evolution, Structure, and Function of Prolamin Storage Proteins. *Seed genomics*. 139-157

Hollingsworth ML, Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM. 2009. Selecting barcoding loci for plants: evaluation of seven candidate loci with species - level sampling in three divergent groups of land plants. *Mol Ecol Res*. 9, 439-457.

Hollingswoth PM, Graham SW, Little DP. 2011. Choosing and using a plant DNA barcode. *PLoS ONE.*6, . e19254.

Hournard NM, Mainville JL, Bonin CP, Huang S, Luethy MH, Malvar TM. 2007. High-lysine corn generated by endosperm-specific suppression of lysine catabolism using RNAi. *Plant Biotechnology Journal.* 5, 605-614.

Hubalkova Z, Rencova E. 2011.One-step multiplex PCR method for the determination of pecan and Brazil nut allergens in food products. J *Sci Food Agric*. 91, 2407-2411.

Huang S, Frizzi A, Florida CA, Kruger DE, Luethy MH. 2006. High lysine and high tryptophan transgenic maize resulting from the reduction of both 19- and 22-kDa zeins. *Plant Molecular Biology*. 61, 525-535.

Hunter BG, Beatty MK, Singletary GW, Hamaker BR, Dilkes BP, Larkins BA, Jung R. 2002. Maize opaque endosperm mutations create extensive changes in patterns of gene expression. *Plant Cell*. 14, 2591-2612.

Jurado-Rivera JA, Vogler AP, Reid CA, Petitpierre E, Gómez-Zurita J. 2009. DNA barcoding insect-host plant associations. *Proc Biol Sci*. 276, 639-948.

Kim KJ, Lee HL. 2004. Complete chloroplast genome sequences from Korean ginseng (Panax shinseng Nees) and comparative analysis of sequence evolution among 17 vascular plants. *DNA Resources*. 11, 247-261.

Kim S. 2008. Processing and properties of gluten/zein composite. *Bioresour Technol*. 6, 2032-2036.

Knowlton N, Weight LA. 1998. New dates and new rates for divergence across the Isthmus of Panama. *Proc. R. Soc. Lond*. 265, 2257-2263.

Kress JM, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH. 2005. Use of DNA barcodes to identify flowering plants. *Proc. Natl. Acad. Sci. U S A*. 102, 8369-8374.

Kress WJ, Erickson DL. 2007. A two-locus rnH-psbA spacer region global DNA barcode for land plants: the coding rbcL gene complements the non-coding. *PLoS ONE*. 2, p. e508.

Krott AA, Caldwell JB, Lilley GG, Higgins TJV. 1991. Amino acid and cDNA sequences of a methionine rich 2S protein from sunflower seed (Helianthus annus L.). *European Journal of Biochemistry*. 195, 329-334.

Lambert RJ, Alexander DE, Dudley JW. 1969. Relative performance of normal and modified protein (*opaque-2*) maize hybrids. *Crop Sci*. 9, 242-243.

Landry J, Moreaux T. 1970. Heterogeneity of corn seed glutelin: selective extraction and amino acid composition of 3 isolate fraction. *Bull. Soc. Chim. Biolo. (Paris).* 52, 1021-1037.

Landry J, Delhaye S, Damerval C. 2000. Improved method for isolating and quantitating alpha-amino nitrogen as non-protein, true protein, salt-soluble proteins, zeins and true glutelins in maize endosperm. *Cereal Chem.* 77, 620-626.

Lang Z, et al. 2004. Cloning of potato SBgLR gene and its intron splicing in transgenic maize. *Plant Science.* 166, 1227-1233.

Larkins BA, Dalby A. 1975. In vitro synthesis of zein-like protein by maize polyribosomes. *Biochem Biophys Res Commun*. 3, 1048-1054.

Larkins BA, Pedersen K, Handa AK, Hurkman WJ, Smith LD. 1979. Synthesis and processing of maize storage proteins in *Xenopus laevis* oocytes. *Proc. Natl. Acad. Sci. U S A.* 12, 6448-6452.

Laskey RA, Earnshaw WC. 1980. Nucleosome assembly. *Nature*. 286, 763-767.

Lending CR, Larkins BA. 1989. Changes in the zein composition of protein bodies during maize endosperm development. *Plant Cell*. 10, 1011-1023.

Li D-Z et al. 2011. Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plant. *Proc Natl Acad Sci U S A.* 108, 19641-19646.

Mcghee JD; Felsenfeld G. 1980. Nucleosome structure. *Annual Rev. Biochem.* 49, 1115-1156.

Mertz ET, Bates LS and Nelson O.E. 1964. Mutant gene that changes protein composition and increases lysine content of maize endosperm. *Science*. 145, 279-280.

Mertz ET. 1976. Genetic improvement of cereal proteins. *Basic Life science*. 7, 465-472.

Nanney DL. 1982. Genes and phenes in tetrahymena. *BioScience*. 32, 783-788.

Negrutiu I, Cattoir-Reynaerts A, Verbruggen I, Jacobs M. 1984. Lysine overproducer mutants with an altered dihydrodipicolinate synthase from protoplast culture of Nicotiana sylvestris (Spegazzini and Comes). *Theor. Appl. Genet*. 68, 11-20.

Organisation, Food and Agriculture. 2003. http://faostat.fao.org/accassed. *FAOSTAT*.

Osborne TB, Mendel LB. 1914. Nutritive properties of proteins of the maize kernel. *The Journal of biological Chemisrty*. 18,1-16.

Osborne TB. 1924. Classification of vegetable proteins. *The vegetables proteins*. 25-35.

Pennisi E. 2007. Taxonomy wanted: a barcode for plants. *Science*. 318, 190-191.

Piccinini S, Negri SA, Lauria M, Viotti A, 2014. Indel and single nucleotide variation of zeins generate unique 2D-zein patterns and molecular markers useful in maize (Zea Mays) genotyping. *Maydica*. 59.

Prasanna BM, Vasal SK, Kassahun B, Singh NN. 2001. Quality protein maize. *Curr Sci*. 81, 1308-1319.

Reyes AR, et al. 2009. Genetic manipulation of lysine catabolism in maize kernels. *Plant Molecular Biology*. 69, 81-89.

Righetti PG, Giannazza E, Viotti A, Soave C. 1977. Heterogeneity of Storage proteins in Maize. *Planta*. 136, 115-123.

Rognes SE, Bright SWJ, Miflin BJ. 1983. Feedback-insensitive aspartate kinase isoenzymes in barley mutants resistant to lysine plus threonine. *Planta*. 157, 32-38.

Rubenstein I, Geraghty DE. 1986. The genetic organization of zein. *In Advances in cereal science technology*. 286-315.

Segal G, Song RT, Messing J. 2003. A new opaque variant of maize by a single dominant RNA-interference-inducing transgene. *Genetics*. 165, 387-397.

Shapiro B, Sibthrope D, Rambaut A et al. 2002. Flight of the dodo. *Science*. 295, 1683.

Sharer TL, Van Oppen MJH, Romano SL, Worheide G. 2002. Slow mitochondria DNA sequence evolution in the Anthozoa (cnidaria). *Molecular Ecology Resources*. 11, 2475-2487.

Shewry PR, Miflin BJ, Bright SWJ. 1981. Conventional and novel approach to the improvement of the nutritional quality of cereal and legume seed. *Sci. Prog.* 67, 575-600.

Soave C, Tardani L, Di Fonzo N, Salamini F. 1981. Zein level in maize endosperm depends on a protein under control of opaque-2 and opaque-6 loci. *Cell*. 27, 403-410.

Song R, Messing J. 2002. Contiguous genomic DNA sequence comprising the 19-kD zein gene family from maize. *Plant Physiol*. 4, 1626-1635.

Spena A, Viotti A, Pirrotta V. 1982. A homologous repetitive block structure underlies the heterogeneity of heavy and light chain zein genes. *The Embo Journal*. 1, 1589-1594.

Sun SSM, Liu QQ. 2004. Transgenic approaches to improve the nutritional quality of plant proteins. *In Vitro Cell. Dev. Biol. Plant.* 40, 155-162.

Tang M, et al. 2013. Nutritional assessment of transgenic lysine-rich maize compared with conventional quality protein maize. *Journal Science Food Agriculture*. 93, 1049-1054.

Torrent M, et al. 1997. Lysine-rich modified gamma-zeins accumulate in protein bodies of transiently transformed maize endosperms. *Plant Molecular Biology*. 34, 139-149.

Tsai, C. Y., and Dalby, A. (1974). Comparisons of the effect of *shrunken-4, opaque-2, opaque-7* and *floury-2* genes on the zein content of maize during endosperm development. *Cereal Chem.* 51, 825-829.

Ufaz S, Galili G. 2007. Improving the content of essential amino acids in crop plants: Goals and opportunities. *Plant Physiology.* 147, 954-961.

Vasal, SK, Villegas E, Bjarnason M, Gelaw B, Goertz P. 1980. Genetic modifiers and breeding strategies in developing hard endosperm opaque-2 materials. In *Improvement of quality traits of maize for grain and silage use,* edited by W.G. Pollmer & R.H. Phillips. Martinus Nijhoff, The Hague. 37-71.

Viotti A, Cairo G, Vitale A, Sala E. 1985. Each zein gene class can produce polypeptides of different size. *The Embo Journal.* 4, 1103-1110.

Viotti A, Pogna NE, Balducci C, Durante M. 1980. Chromosomal localization of zein genes by in situ hybridization in *Zea mays*. *Mol. Gen. Genet*. 1, 35-41.

Viotti A, Sala E, Alberi P and Soave C. 1978. Heterogeneity of zein synthesized in *vitro*. *Plant Science Letters.* 13, 365-375.

Viotti A, Sala E, Marotta R, Alberi P, Balducci C, Soave C. 1979. Genes and mRNAs coding for zein polypeptides in Zea mays. *European Journal of Biochemistry*. 102, 211-222.

WHO. 2006. Nutrition for Healthy and Development.

WHO/FAO/UNU. 1985. Energy and protein requirements.

Wilson. 1983. Seed protein fraction of maize, sorghum and related cereals. *Seed proteins: biochemistry, genetics, nutritive value.* 271-307.

Woo YM, Hu DW, Larkins BA, Jung R. 2001. Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. *Plant Cell*. 13, 2297-2317.

Yu J, et al. 2004. Seed-specific expression of the lysine-rich protein gene sb401 significantly increases both lysine and total protein content. *Molecular Breeding.* 14, 1-7.

Yue J, Li C, Zhao Q, Zhu D, Yu J. 2014. Seed-Specific Expression of a Lysine-Rich Protein Gene, GhLRP, from Cotton Significantly Increases the Lysine Content in Maize Seeds. *International Journal of Molecular Science.* 15, 5350-5365.

Zhang DX, FM Hewitt. 1997. Insect mitochondrial control region: a review of its structure, evolution and usefulness in evolutionary studies. Biochem. *Syst. Ecol*. 25, 99-120.

Acknowledgements

I would like to express my sincere gratitude to my tutor Dr. Angelo Viotti for his patience, motivation, and huge knowledge. His guidance helped me through all the research and writing of this thesis.

I would like to thank Prof. Giovanni Dehò for the opportunity he gave me. Without his help I would not have reached my final goal.

I would like to thank my friends Dr. Massimiliano Lauria and Dr.ssa Franca Locatelli for their continuous help and support in my PhD study and research.

I would also like to thank my parents and parents in law, for their loving support and encouragement.

Finally, I would like to thank my husband Paolo who has always been cheering me on and standing by me through the good and bad times.

PART II

Published paper_1 in *Maydica*

Original Paper Open Access

Indel and single nucleotide variations of zeins generate unique 2D-zein patterns and molecular markers useful in maize (*Zea mays*) genotyping

Sara Piccinini[1], Alfonso Simone Negri[2], Massimiliano Lauria[1], Angelo Viotti[1]*

[1]Istituto di Biologia e Biotecnologia Agraria, CNR, Via Bassini 15, 20133 Milano, Italy
[2]Dipartimento di Scienze Agrarie ed Ambientali, Università di Milano, Via Celoria 2, 20133 Milano, Italy
*Corresponding author: E-mail: viotti@ibba.cnr.it

Abstract

In this study, we investigated the inter- and intra-genomic sequence variation of alpha-zein genes and their polypeptide expression in different maize genotypes, i.e. inbreds and a set of Lombardy open pollinated varieties, by analyzing their RFLP, coding nucleotides and 2-dimensional (2D) protein fractionation profiles. An extensive analysis of coding capacity of alpha-zein sequences in various genotypes and in the B73 reference inbred allowed us to assign 2D-spots to specific zein sequences. Moreover, we found that some genes reported to contain in frame stop codons are very likely expressed. Collectively these data allowed us to constitute two barcodes respectively based on nucleotide variation and on 2-D protein patterns that identify univocally each genotype.

Keywords: zein gene RFLP, 2D zein polypeptides, stop codons, barcoding, genotyping

Introduction

In the maize kernel, storage protein genes are expressed only in the maize endosperm where their products are exclusively accumulated during grain filling. Those genes encoding for those polypeptides belong to two biochemical classes, i.e. glutelin and prolamine, differing for their peculiar solubility in alcoholic solvents in presence or absence of S-S bridges reducing agent (Soave et al, 1975; Landry and Moureaux, 1980). Glutelins were further subdivided in G1L plus G1S, G2 and G3 sub-classes currently termed gamma- beta- and delta-zeins, respectively, while zeins in sensu stricto i.e. the prolamine class was termed alpha-zein (see Lopes and Larkins, 1998, for a review).

The alpha-zeins were one of the first cloned and molecularly characterized genes and polypeptides in plants (Righetti et al, 1977; Burr et al, 1982; Rubenstein et al, 1982; Viotti et al, 1982. Furthermore, data on nucleotide hybridization and polypeptide analyses clearly indicated the occurrence of a multigene family divided into various subfamilies. (Righetti et al, 1977; Viotti et al, 1978, 1979). Additional stidies about their accumulation in sub-cellular storage organelles surrounded by membranes of rough endoplasmic reticulum (RER) suggested in the zein polypeptides the presence of a signal peptide responsible for their targeting to the endomembranes (Burr et al, 1978). This observation was further confirmed by nucleotide sequence and amino acid comparison analyses to occur at the amino terminus with a constant length of 21 amino acids for all the alpha-zeins (Nien-Tai et al, 1982; Viotti et al, 1982). Similarly, the coding capacities of the mature alpha-zeins were shown to range between 212 and 246 amino acids (Burr et al, 1982; Spena et al, 1983; Larkins et al; 1985, Viotti et al, 1985).

Their wide spectrum in size, charge and their different relative abundance of zeins among genotypes, have indicated a complex system of transcriptional/expression control in time and space (Dolfini et al, 1992; Ciceri et al, 2000; Song and Messing, 2003). Furthermore, detailed nucleotide analyses of both genomic and cDNA sequences of inbreds also revealed that the zein genes are clustered in few chromosomal regions and that a collinear relationships did not always exist among genotypes (Song and Messing, 2002, 2003). In addition, it was found that the occurrence of single nucleotide variations and in frame short- or long-indel within each sub-group modify their coding capacities and charge heterogeneity. In this context, it was also noted that the nucleotide sequence identity, rather than the length in their coding capacity, had permitted to classify zeins into four subfamilies SF1, SF2, SF3, and SF4 (Larkins et al, 1985; Viotti et al, 1985). Among these subfamilies the latter one was the most abundant with about 20 copies per haploid genome (Song and Messing, 2003). However, because of partial chromosomal duplication events within the alpha-zein cluster, the total copy number of each subfamily was shown to vary among inbreds (Song and Messing, 2003). Studies, by comparing genes and alleles among inbreds and collecting data on cDNA libraries, indicated that the zein system shows the occurrence of in frame prestop codons in sequences from genomic fragments (Song and Messing, 2002, 2003). Therefore, this amplifies, by consequence, the heterogeneity of the

RECEIVED 07/06/2014

nucleotide sequences, the polypeptide expression and the 2D zein patterns. Accordingly, in this study we have investigated the zein sequences and polypeptides heterogeneities among a set of maize inbred lines and open pollinated varieties of maize. Our results indicated that the zein system is a valuable tool useful for maize genotype identification in maize. In addition, the analyses on the 2D fractionations of alpha-zein polypeptides compared to genomic sequences of two elite inbred lines, indicate that some genes characterized by the presence of stop codons in the coding sequence are expressed.

Materials and Methods

Plant Material and Growth Conditions

For the different genotypes used in this study, plants were grown in the field or in the greenhouse during 2009 and 2010. Few plants of each genotype were used to obtain immature ears, 4-5 cm long, to extract DNA. Seeds were harvested at maturity. A list of the genotypes used in this study is reported in Supplementary Tables 2 and 3.

Nucleic Acid Isolation and DNA Blot Analysis

DNA was extracted and purified as described (Dolfini et al, 1992; Bernard et al, 1994). Purified DNA was digested to completion (Bianchi and Viotti, 1986) and DNA gel blot analysis were carried out as previously reported (Bernard et al, 1994). Fractionation of digested DNAs was in 0.8-1.0% agarose gel in TAE buffer. Probes were labelled with the Rediprime Kit according to the manufacturer procedure (GE Healthcare) by addition of [α-32P]dCTP. The zein cDNA (M6, E19, and M1) probes were those already described (Viotti et al, 1985; Dolfini et al, 1992). The D3 cDNA probe was obtained from cloning and analyses of cDNA from endosperm RNA of W64a as given by Viotti et al (1982). Hybridization of the various genotypes occurred with either the prototype sequences or the amplicons of the different subfamilies obtained and characterized as described in the following section. Washing conditions of the nucleic acid filters were performed at high stringency (Viotti et al, 1982, 1985), which discriminates identity lower than 96%. Filters were exposed to X-ray film with or without intensifying screens or by their scanning with Starion (Elite Healthcare).

DNA amplification and analysis of amplicon specificity

Amplified zein sequences of the four subfamilies, i.e. SF1, SF2, SF3, and SF4, were obtained using DNA of the W64a inbred and purified as described above. Specific primers of each subfamily are reported in Supplementary Table 1. Fractionation occurred in 2.0 NuSeive agarose gel (FMC) in TBE buffer. The amplified fragments were checked for their group identity using digestion of part of the total amplified fragments with specific restriction enzymes. For instance, all the SF4 have only one *HincII* site that is absent in all the other three subfamilies. Similarly, all the SF2 have only one *BamHI* site absent in all the other three subfamilies. For SF1 or SF3 we used *PstI* and *VspI*, respectively. In all instances, we made a quadruple check of digestions for each subfamily with the above-mentioned restriction enzymes to ascertain presence or absence of digestion or even generation of different fragment sizes, as it occurs for the SF1 and SF4 when using *PstI* that is present only in some SF4 sequences while occurring in all the SF1.

PCR polymorphism of B73 accessions

Four forward primers spanning the 5' coding re-

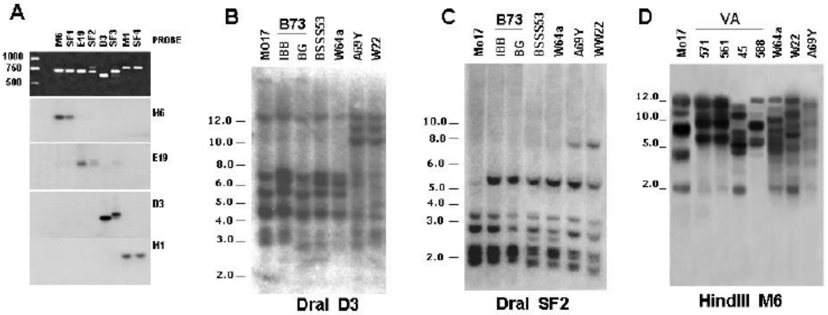


Figure 1 - Analysis of probe specificity and RFLP.

A: Four gels were run in parallel. Each gel was loaded with the four prototype sequences M6, E19, D3, and M1 together with the amplified fragments that identify each subfamily, respectively, SF1, SF2, SF3, and SF4, as indicated on the top. DNAs were transferred to filters and each filter was hybridized to probes as indicated to the right. Markers in bp are reported on the left. **B**, **C**, and **D**: DNA gel blot analysis of genomic DNA of different genotypes digested with different restriction enzymes and probed as indicated at the bottom of each panel. Varieties are indicated with their number as reported in Supplementary Table 1. Size markers in kb are indicated on the left.

gion of different alpha-zein genes were constructed together with two as reverse primers designed in the mature coding part considering published sequences of both the B73 and BSSS53 inbreds (Woo et al 2001; Song and Messing, 2003). They are reported in Supplementary Table 1. Two Taq systems were used: GoTaq (M7112, Promega) and 5' Hot start (5Prime). In both a final concentration of 2.5 mM Mg ion was used. PCR reactions were carried out in the presence of 2% deionised formamide with annealing temperature at 58°C for 33 cycles. Other conditions were according to manufacturer protocols. Fragments were resolved in 2.2% Metaphore agarose gel (FMC) in TBE buffer.

Storage protein extraction and electrophoresis analysis

Extraction of glutelin and prolamine from mature seeds (five-six per each genotype from the central part of the ear) was carried out as described in Lund et al, (1995) in the presence of 2% beta-mercapto-ethanol in 70% ethanolic solution. Mono-dimensional (1D) fractionation by SDS-PAGE and two-dimensional (2D) by IEF/SDS-PAGE were carried out essentially as described (Lund et al, 1995). IEF markers used in 2D fractionation were purchased from GE-Health-care. Gels were stained with Comassie Brilliant Blue R250.

Nucleic acid and protein Barcodes

The obtained DNA banding pattern for presence or absence of each fragment was converted in a binary code and then into a barcode using the free online software barcode.tec-it/com. The 2D gels were scanned in an Epson Expression 1680 Pro Scanner and analyzed with ImageMaster 2-D Platinum Software v6.0 (GE Healthcare Life Sciences, USA). Automatic matching was complemented by manual matching. Molecular weights of the spots were estimated using the migration range standard as reported in Viotti et al (1982), while pI was determined as described in Righetti et al (1977).

Results

Zein probe recovery and specificity

In our analysis we refer to zein prototypes the M6, E19 and M1cDNA sequences, belonging to alpha-zein genes encoding for polypeptides of the subfamilies SF1, SF2 and SF4, respectively. (Viotti et al, 1982, 1985). Recently, during a screening of clones

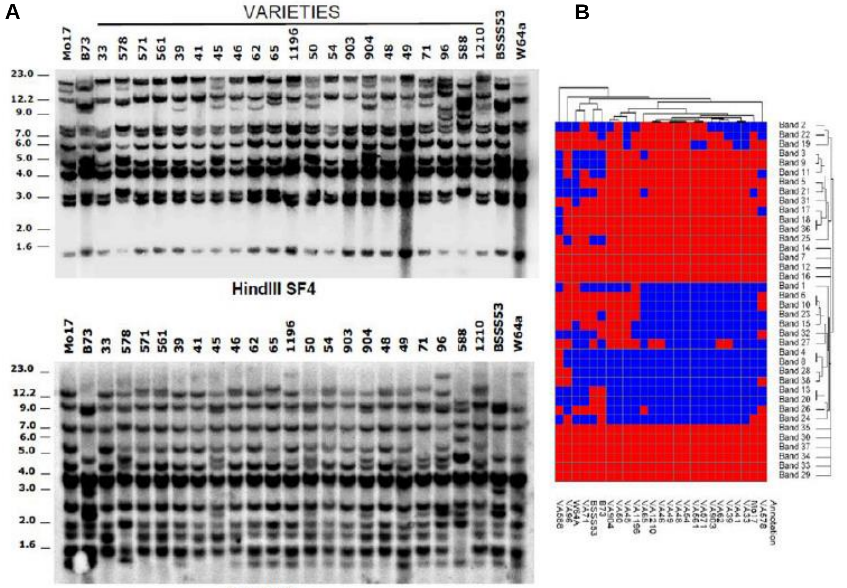


Figure 2 - RFLP of maize varieties and inbreds.

A: DNA gel blot analysis of genomic DNA of the varieties and four inbreds as reference. DNAs were digested and probed as indicated at the bottom. Varieties are indicated with their number as reported in Supplementary Table 1. Size markers in kb are indicated on the left. B: Heatmap of the presence/absence of the RFLP bands obtained from the blots in A.

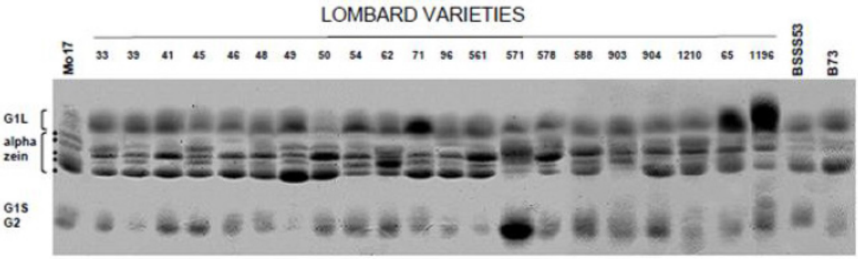


Figure 3 - Mono-dimensional (1D) fractionation of ethanol extracts.
Zeins were from mature seeds of maize varieties and of three inbreds. On the left relative mobility of mature gamma-zein (long type, G1L), alpha-zein, gamma-zein (short type, G1S), and G2 are indicated. Varieties are indicated with their number as reported in Supplementary Table 1. Dots on the left indicate the position of the five major size classes of alpha-zeins.

from the cDNA library of the W64A inbred (Viotti et al, 1982), we identified a clone, D3, belonging to the SF3. This clone and a cDNA clone from the inbred Ohio43 (GeneBank accession AI677029.1) represent variants of the SF3 subgroup, containing a deletion of 93 nucleotides when compared with other sequences of the same subfamily (Supplementary Figure1). Apart from the internal deletion, D3 has an identity of 99.6% to the az19D1 cDNA (GeneBank accession NM001111586.1).

We designed SF-specific primers for each zein prototype at the 5' and at the 3' ends of the coding sequences (Supplementary Table1) and used them to amplify, form genomic DNA, PCR products that represent the entire set of sequences of each subfamily. Amplification results are reported in Figure 1A together with the results of their hybridization specificity as determined by DNA gel blot analysis. The specificity and uniqueness of sequences is further supported by results of specific restriction enzyme analysis as reported in Materials and Methods.

Genomic analysis of alpha zein genes in inbreds and varieties

The PCR products of each SF together with the four zein prototypes were used as probes in DNA blot experiments aimed to characterize the extent of DNA polymorphisms at the alpha zein loci among a set of Lombard open pollinated varieties. A preliminary analysis was performed on a subset of these varieties and elite inbred lines, representing different maize breeding group, to identify the more informative probe-restriction enzyme combination for fingerprinting purpose. The list of the various genotypes is reported in Supplementary Tables 2 and 3. DNAs was extracted from sporophytic tissue and digested with the *HindIII* and *DraI* restriction enzymes. These enzymes were chosen for the analysis as they do not digest DNA within the zein coding region and are insensitive to DNA methylation (Bianchi and Viotti, 1988). Although each probe-restriction enzymes combination produces DNA blot patterns that are specific for each zein SF (Figure 1B, C, D and data not shown), we found that only the SF4-HindIII and SF4-DraI combinations were the more informative (Figure 2A). Indeed, the analysis of the entire set of Lombard varieties with these two probe-restriction enzymes combinations permitted to generate on the basis of presence/absence of the various bands, a binary code, this was transformed into a barcode that identified univocally the twenty-five genotypes investigated herein (Supplementary Table 4). The heatmap diagram depicts the uniqueness and creates a relationship among the various genotypes (Figure 2B).

Zein polypeptide analysis by 1D and 2D fractionations

The set of the Lombard varieties and three or four inbred lines were fractionated by mono-dimensional (1D) (SDS-PAGE) and two-dimensional (2D, IEF/SDS-PAGE) analyses. As expected on the basis of the previous results (Viotti et al, 1982; Ciceri et al, 2000) each genotype showed different and specific 1D pattern of the five size classes of alpha-zein (Figure 3). Some genotype as Va46 misses the H2 and few other genotypes some bands in the light size classes. The G1L or gamma-zein-27 migrates as a broad band in 1D and as a number of smeared spots in 2D (Figures 3, 4A and Supplementary Figure 2). The two other glutelins, G1S and G2, migrate faster and generate different spots. The fractionations of the Lombard varieties show that none of the 1D patterns (Figure 3) or the 2D patterns (Figure 4A and Supplementary Figure 2) have either identical relative abundance of the five size classes or 2D-spots distribution, with a number of spots ranging from eight to thirteen for VA54 and W64a, respectively.

In the 2D analysis five IEF markers were properly chosen and added to each extract (roman numbers). The 2D spot distribution of each genotype was compared to that of the W64a considering the position of the five IEF markers and that of the two spots of G1S (A and B as in the panels of Figure 4A and Supplementary Figure 2). The software program, as reported

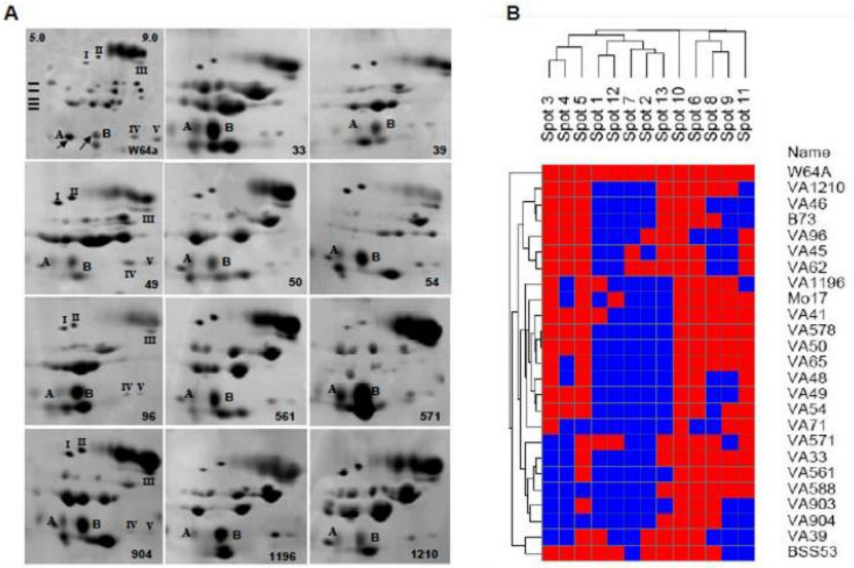


Figure 4 - Two-dimensional (2D) fractionation of storage protein polypeptides.
A: Total ethanol extracts from mature seeds of maize varieties and the W64a inbred were subjected to isoelectrofocusing in the first dimension and then to SDS-PAGE in the second dimension. Loading was about 130 μg for each genotype. The pH ranges are reported on the top of the upper left panel and apply for all the panels. Roman numbers in the four left panels identify IEF markers. In each panel the position of the two polypeptides of the short gamma-zein (G1S) is also indicated (A and B) that together with the IEF markers are utilised for proper positioning of the alpha-zein spots. Dashes in the upper right panel identifies from top to bottom the five major size classes of alpha-zein polypeptides, H1, H2, L1, L2 and L3, respectively. **B**: Heatmap of the 2D fractionations of the twenty-five genotypes reported in **A** and Supplementary Figure 3.

in Materials and Methods, applied to the twenty-five panels of the two figures allowed us to obtain a binary code (Supplementary Table 5) and a cluster reported in Figure 4B. As for the DNA RFLP each genotype can be univocally identified.

Correlation of 2D spots to sequenced alpha-zein genes of heavy type and SNP analysis in B73 accessions

This broad variation in 2D patterns prompted us to verify in different accessions and year harvesting of B73 the expression stability of zein genes. We considered eight different accessions of B73 inbred line because zein sequences of this inbred were reported from different laboratories (Woo et al, 2002; Song and Messing, 2003). Moreover, for this line a high quality reference genome is available. DNA gel blot analyses (Figure 5A) of the eight accession yielded identical patterns, with the exception of the BG accession that showed in the HindIII-D3 combination an additional band of about 18 kb and a missing band around 4 kb. Similarly, the same accession showed an additional band of about 10.5 kb in the HindIII-SF2 combination. However, in the HindIII-SF4 combination the RFLP was identical for all the eight accessions (data not shown).

2D analysis of the eight accessions revealed a constant pattern in term of spot distribution of the H1 and H2 size classes, although, some differences in their intensities were observed. Six spots are present in the H1 region and three spots in the H2 region (Figure 5B). This pattern was compared to that of the BSSS53 inbred, whose entire gene-cluster coding for H1 and H2 was sequenced (Song and Messing, 2003). The spots of the two inbreds were compared for their relative position and for the identification of genes responsible for their expression (Table 1). Some of those genes were reported to contain in frame stop codons (Song and Messing, 2003), however, their substitution with a coding codon generates polypeptides that are present with a good correspondence in the 2D patter, moreover, the intact gene 22.12 does not generate a transcript in the reported expression analysis (Song and Messing, 2003). For instance, the 22.20 gene present in both the inbreds BSSS53 and B73 corresponds as coding characteristics to spot -1 in the former and spot 1 in the latter (Figure 5B

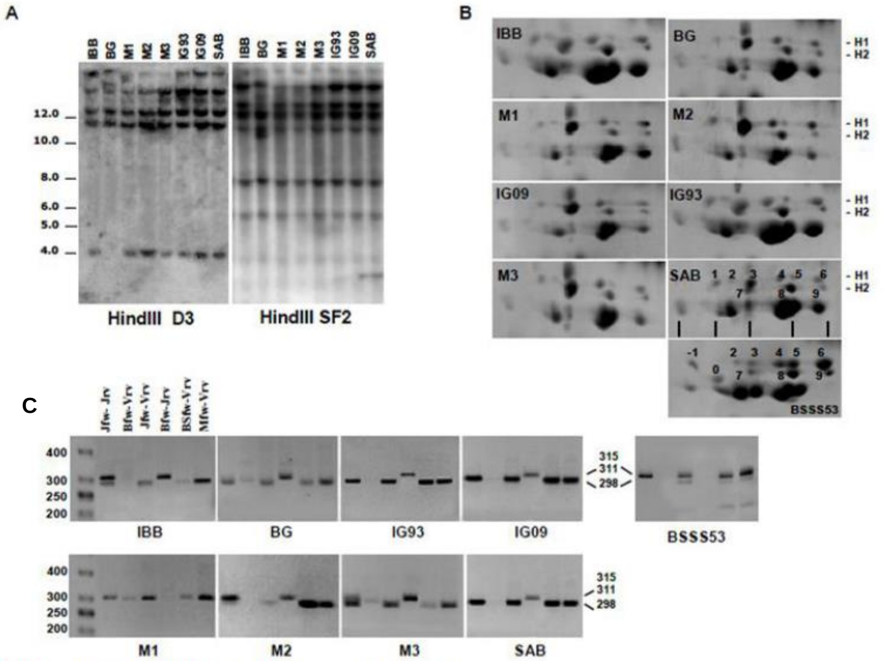


Figure 5 - Genomic DNA and alpha-zein analyses of B73 accessions.

A: Genomic DNA of the eight accessions was digested and probed as indicated at the bottom of each panel. Size markers in kb are indicated on the left. **B**: Fractionation by 2D of alpha-zein of the eight B73 accessions and of the Bsss53 as comparison. The two size classes of heavy type of alpha-zein are indicated on the right. In the SAB and BSSS53 panels the spots of H1 and H2 are numbered. These two panels were used for comparison and correlation analyses of spots and coding capacity of genes, see Table 1. In SAB panel the five vertical bars indicate from left to right pH 5.0, 6.0, 7.0, 8.0, and 9.0, respectively. **C**: The amplified fragment from the B73 accessions and BSSS53 was obtained and fractionated as reported in Materials and Methods. On the upper left panel combinations of primers, as listed in Supplementary Table 1, are indicated and apply for all the panels. On the right or on the left of the panels size markers are indicated in bp. The abbreviated name of the eight B73 accessions is reported in SupplementaryTable 3.

and Table 1, see note and comments). In analyzing the entire set of genes we succeeded to assign most of the spots to the published sequences. This type of analysis indicates: i) that the substitution of single base in generating in frame stop codons may have derived during bacterial cloning or ii) the absence of transcript for intact genes or alleles, as in the case of the *22.12* (spot 2 in both inbreds), may reflect mRNA paucity at the developmental stage tested, or even error in their identification (Woo et al, 2001; Song and Messing 2003; Luo et al, 2008; Feng et al, 2009).

In comparing alpha-zein sequenced genes of B73 present in the MaizeGDB and NCBI data Banks we noticed several nucleotide variation of the same allele. For instance the *floury2* locus has been reported to code for a mature zein polypeptide of 241 aa (Woo et al, 2001) or 242 aa (Song and Messing, 2003; MaizeGDB). Similarly, the *az22z1* gene (GenBank Accession AF371274; Woo et al, 2001) codes for a mature polypeptide of 242 aa, a unique sequence among the other annotated genes of the two B73 (Song and Messing, 2003; MaizeGDB). Comparison of the 5' nucleotide end of their coding region showed several SNPs for the same allele from the various data banks and even between alleles of the B73 and BSSS53 lines, which allowed the construction of different primers (Supplementary Table 1) with specific SNPs used in PCR reaction and challenged against the DNAs of the eight B73 accessions (Figure 5C). As expected, the patterns of amplification of the eight accessions had two main bands, one around 311-315 nt and the other at 298 nt, respectively. By comparison also the BSSS53 was considered. This line manifested an additional band of 227 nt that derives from the amplification of the *Zp22/D87* gene, absent in B73 (Song and Messing, 2003), and occured only with two combinations of primers. In comparing the patterns of the eight B73 and starting from the left couple of primers (Figure 5C) it was noted: i) the first couple generates in some cases a double band (IBB, M2 and M3) or a single band (the fast migrating) in all the others; ii) the second couple a faint band in BG and M3, or its absence in all the others apart M1; iii) in M1 only the upper band in all the primer combination; iv) the last four couples of primers generate an identical pattern for the eight B73 apart M1.

Discussion

Our data from RFLP and 2D fractionation assayes, with the generation of the corresponding barcode, identify each genotype as unique. These findings indicate that the alpha-zein system has acquired during generations a large heterogeneity generating a broad diversification and therefore yelding specific haplotype (Song and Messing, 2003). The D3 clone represents such an example as it is recovered into different inbreds, namely W64a and Ohio43. A further indication of the broad diversification occurring among genotypes derives from the analysis of two cDNA clones azs1-42 and azs2-2 (GeneBank accession AJ491308 and AJ491309, respectively) that were previously identified in the NYRo2 inbred line (Ciceri et al, 2000). These sequences have a 98% identity to those of the BSSS53 Zp22/6 and Zp22/D87, respectively. The azs1-42 produces a polypeptide in the position of spot 6 (as Zp22/6 does, Figure 5B) and the azs2-2 a spot toward more acidic pH, pI to 5.44, in respect of Zp22/D87 pI 6.74, because azs2-2 contains the aa substitution 42A>E (Lauria and Viotti, unpublished). This single-nucleotide mutation, as for B73 and BSSS53 genes, may derive from the proposed mutational mechanism associated to insertions/deletions occurring close to genes (Tian et al, 2008; Hollister et al, 2010), as was showed for absence of collinearity and frequent indels in zein gene clusters (Song and Messing, 2002, 2003).

The dendograms of the RFLP bands and 2D patterns (Figures 4B and 5B) generate a different relationship among genotypes because they are based onto different types of data that clearly discriminate zein alleles: inter genic variation and gene expression. In the eight accessions of B73 the inter genic sequences, at least for the three restriction enzymes used (Figure 4A, for *EcoRI* data not shown), were almost consistent in their RFLPs, whereas differences were observed in the coding sequences by PCR analysis (Figure 5C). This suggests that zein intra genic sequences may vary with higher frequency in respect of inter genic regions generating transcript isoforms specific for each genotype. Accordingly, 2D fraction-

Table 1 - Correlation of 2D alpha zein spots and gene coding capacity and characteristics.

BSSS53 spot*	pI/MW[&]	gene/ transcript*	note	B73 spot*	pI/MW[&]	gene/ transcript*	note[$]
-1	5.27/26740	22.20/no	stop[a]	1	5.99/26740	22.20/no[a]	stop[a]
2	6.91/26709	22.12/no	intact[b]	2	6.91/26709	22.12/no	intact[b]
3	7.01/26852	22.21/no	stop	3	7.01/26760	22.9/yes	intact
4		NI[c]		4	8.09/26877	22.8/yes	intact
5	8.11/26891 8.14/26700	22.8/yes 22.10/yes	intact intact	5	8.11/26838	22.19/yes	intact
6	~8.94/~26800	22.14+22.4 +Zp22/6[e]/yes	intact	6	8.95/26760	22.4/yes	intact
7	7.01/~26500	NI		7	7.01/~26500	NI	
8	8.11/26358	fl2/yes	intact	8[d]	8.11/26358	fl2/yes	intact
9	8.95/26500	NI		9	8.95/26500	NI	
10	6.74/23527	Zp22/D87[e]/yes	intact		//	Zp22/D87[e]	absent

*The gene name and the transcript occurrence were derived from Song and Messing, 2003. It should be noted that transcript level was carried out on endosperm tissue at 18 DAP, while spots are from mature endosperm and are the result of an accumulation process. On the other hand transcript level does not necessarily reflect a proportional amount of polypeptide as translational control of alpha-zein mRNA has been reported Spena et al (1985).

[&] The theoretical Isoelectric point (pI) and the MW were calculated with ExPASy program (http://web.expasy.org/compute_pi/)

[$] Reports the annotations of Song and Messing (2003).

[a] The correction of the stop codons in the BSSS53 and B73 alleles generate in the former a polypeptides with a pI of 5.27 that in respect of the 22.20 allele of B73 contains an aa substitution 157Q>E that modifies its focalization towards more acidic pH (spot -1).

[b] The two alleles are intact, but their transcripts were not detected in the reported analysis (Song and Messing, 2003). In any case their theoretical pI and MW fit spot 2.

[c] Not Identified.

[d] The gene *az22z1* of the B73 reported by Woo et al (2001) has a pI of 8.13 and MW of 26358 that fit spot 8. The other four genes reported in this article, az22-3, -4, -5 and –fl2 fit spots, 6, 3, 2 and 8, respectively (data not shown).

[e] Genes absent in the B73. The Zp22/D87 in BSSS53 contains an internal deletion that generates the 227 nt band reported in PCR analysis of Figure 5B.

ation (Figure 4A and Supplementary Figure 2) of the twenty-five genotypes showed a number of alpha-zein spots that vary from about eight up to thirteen, with the H1 and H2 size classes showing the most heterogeneous pattern in respect of the group of light size classes. Moreover, so far none of the 2D patterns presently analyzed in our laboratory for more than thirty-five genotypes were shown to be identical to each others (present data, Lund et al, 1995; Ciceri et al, 2000). In addition, the 2D analysis indicates that single aa substitution that involves neutral to charge aa, as in the case of the alleles of the 22.20 gene of BSSS53 and B73, results into a distinct spot position (Figure 5B and Table 1). This is also the case for the genes *Zp22/D87* and *azs2-2* as discussed above.

A further consideration suggested us that some spots of major intensity, as spot 6 in BSSS53 in respect of the same spot in B73, may reflect the expression of three genes versus only one in B73. Moreover, on the basis of notes and considerations reported in Table 1, spot 3 in the BSSS53 and B73 appeared result of the expression of two different genes, *22.21* in the former and *22.9* in the latter: note that the *22.21* is absent in the cluster of B73 and the only one that fit spot 3 has to be the *22.9*. Moreover, the 22.9 sequence of BSSS53 has several stop codons and a short rearrangement at the 3' end that strongly may hamper its expression as polypeptide. On the other hand, some genes of both B73 and BSSS53 were reported to contain only one in frame stop codon that always involve the transition C>T of the two triplets CAA and CAG coding for glutamine (Song and Messing, 2002, 2003). In this analysis we have found that some of them are expressed indicating that base transition may occur during cloning and amplification in the bacterial cell. Collectively, alleles may generate two spots with different pI, but on the other hand one spot of different genotypes with the same pI may be the result of the expression of two or more different genes.

The overall data from the current study clearly indicate that among genotypes the entire set of genes of the alpha zeins has a high degree of variation in the coding region and that even genotypes can accumulate dominant negative mutations (Coleman et al, 1995; Kim et al, 2004; Kim et al, 2006) for these polypeptides important for feed utilization that are, however, not essential for maize reproduction.

Acknowledgements

We thank Dr Mario Motto for helpful comments on the manuscript, Dr Raul Pirona for help in manuscript editing and the MGCSC for providing seed stocks. This research was supported by Regione Lombardia (Italy), agreement Regione/CNR, project 5: ''Nuovi strumenti per la selezione di varietà di mais con migliore valore proteico''.

References

Burr B, Burr FA, Rubenstein I, Simon MN, 1978. Purification and translation of zein messenger RNA from maize endosperm protein bodies. Proc Natl Acad Sci USA 75: 696-700

Burr B, Burr FA, St John TP, Thomas M, Davis RW, 1982. Zein storage protein gene family of maize. An assessment of heterogeneity with cloned messenger RNA sequences. J Mol Biol 154: 33-49

Burr FA, Burr B, 1981. In vitro uptake and processing of prezein and other maize preproteins by maize membranes. J Cell Biol 90: 427-434

Burr FA, Burr B, 1982. Three mutations in Zea mays affecting zein accumulation: a comparison of zein polypeptides, in vitro synthesis and processing, mRNA levels, and genomic organization. J Cell Biol 94: 201-206

Ciceri P, Castelli S, Lauria M, Lazzari B, Genga A, Bernard L, Sturaro M, Viotti A, 2000. Specific combinations of zein genes and genetic backgrounds influence the transcription of the heavy-chain zein genes in maize *opaque-2* endosperms. Plant Physiol 124: 451-460

Clark RM, Linton E, Messing J, Doebley JF, 2004. Pattern of diversity in the genomic region near the maize domestication gene *tb1*. Proc Natl Acad Sci USA 101: 700-707

Coleman CE, Clore AM, Ranch JP, Higgins R, Lopes MA, Larkins BA, 1997. Expression of a mutant alpha-zein creates the floury2 phenotype in transgenic maize. Proc Natl Acad Sci USA 94: 7094-7097

Dolfini S, Landoni M, Tonelli C, Bernard L, Viotti A, 1992. Spatial regulation in the expression of structural and regulatory storage-protein genes in Zea mays endosperm. Dev Gen 13: 264-276

Geraghty DE, Messing J, Rubenstein I, 1982. Sequence analysis and comparison of cDNAs of the zein multigene family. EMBO J 1: 1329-1335

Kim CS, Gibbon BC, Gillikin JW, Larkins BA, Boston RS, Jung R, 2006. The maize Mucronate mutation is a deletion in the 16-kDa gamma-zein gene that induces the unfolded protein response. Plant J 48: 440-451

Kim CS, Hunter BG, Kraft J, Boston RS, Yans S, Jung R, Larkins BA, 2004. A defective signal peptide in a 19-kD alpha-zein protein causes the unfolded protein response and an opaque endosperm phenotype in the maize *De*-B30* mutant. Plant Physiol 134: 380-387

Kirihara JA, Petri JB, Messing J, 1988. Isolation and sequence of a gene encoding a methionine-rich 10-kDa zein protein from maize. Gene 71: 359-70

Landry J, Moureaux T, 1980. Distribution and Amino-Acid-Composition of Protein-Fractions in Opaque-2 Maize Grains. J Agric Food Chemistry 21: 1865-1869

Lopes MA and Larkins BA, 1993. Endosperm Origin, development, and function. Plant Cell 5: 1383-

1399

Lund G, Ciceri P, Viotti A, 1995. Maternal-specific demethylation and expression of specific alleles of zein genes in the endosperm of *Zea mays* L. Plant J 8: 571-581

Luo M, Liu J, Lee RD, Guo BZ, 2008. Characterization of gene expression profiles in developing kernels of maize (*Zea mays*) inbred Tex6. Plant Breeding 127: 569-578

Marks MD, Lindell JS, Larkins BA, 1985. Nucleotide sequence analysis of zein mRNAs from maize endosperm. J Biol Chem 260: 16451-16459

Nien-Tai H, Peifer MA, Hidecker G, Messing J, Rubenstein I, 1982. Primary structure of a genomic zein sequence of maize. EMBO J 1: 1337-1342

Righetti PG, Gianazza E, Viotti A, Soave C, 1977. Heterogeneity of storage proteins in maize. Planta 136: 115-123

Soave C, Pioli F, Viotti A, Salamini F, Righetti PG, 1975. Synthesis and heterogeneity of endosperm proteins in normal and *opaque-2* maize. Maydica XX: 83-94

Soave C, Suman N, Viotti A, Salamini F, 1978. Linkage relationships between regulatory and structural gene loci involved in zein synthesis in maize. Theor Appl Genet 52: 263-267

Soderlund C, Descour A, Kudrna D, Bomhoff M, Boyd L, Currie J, Angelova A, Collura K, Wissotski M, Ashley E, Morrow D, Fernandes J, Walbot V, Yu Y, 2009. Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs. PLoS Genet 5: e1000740

Song R, Llaca V, Linton E, Messing J, 2001. Sequence, regulation, and evolution of the maize 22-kD alpha zein gene family. Genome Res 11: 1817-1825

Song R, Messing J, 2002. Contiguous genomic DNA sequence comprising the 19-kD zein gene family from maize. Plant Physiol 130: 1626-1635

Song R, Messing J, 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. Proc Natl Acad Sci U S A 100: 9055-9060

Spena A, Krause E, Dobberstein B, 1985. Translation efficiency of zein mRNA is reduced by hybrid formation between the 5'- and 3'-untranslated region. EMBO J 4: 2153-2158

Spena A, Viotti A, Pirrotta V, 1983. Two adjacent genomic zein sequences: structure, organization and tissue-specific restriction pattern. J Mol Biol 169: 799-811

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, Hudson R, Bergelson J, Chen JQ, 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. Nature 455: 105-108

Viotti A, Balducci C, Weil JH, 1978. Adaptation of the tRNA population of maize endosperm for zein synthesis. Biochim Biophys Acta 517: 125-132

Viotti A, Cairo G, Vitale A, Sala E, 1985. Each zein gene class can produce polypeptides of different sizes. Embo J 4: 1103-1110

Viotti A, Sala E, Marotta R, Alberi P, Balducci C, Soave C, 1979. Genes and mRNAs coding for zein polypeptides in *Zea mays*. Eur J Biochem 102: 211-222

Woo YM, Hu DW, Larkins BA, Jung R, 2001. Genomics analysis of genes expressed in maize endosperm identifies novel seed proteins and clarifies patterns of zein gene expression. Plant Cell 13: 2297-2317

Wu Y, Goettel W, Messing J, 2009. Non-Mendelian regulation and allelic variation of methionine-rich delta-zein genes in maize. Theor Appl Genet 119: 721-731

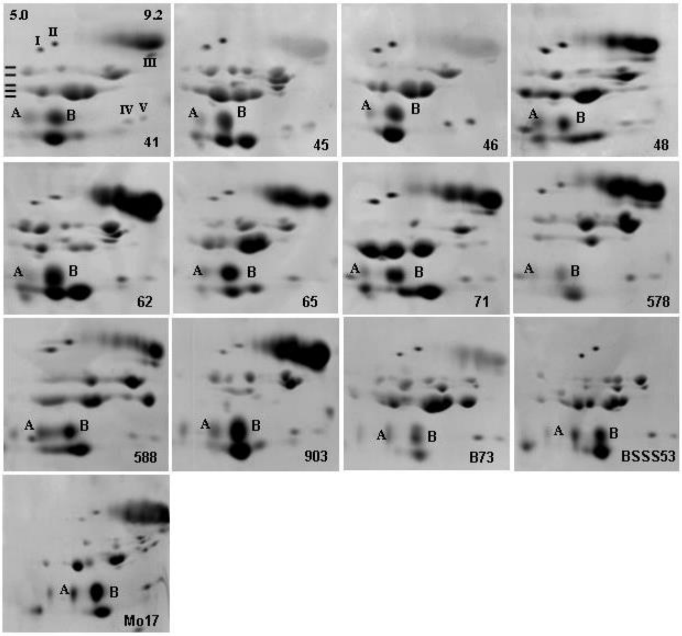
Supplementary data of paper

S_figure 1- Sequence of clone D3 and main features.

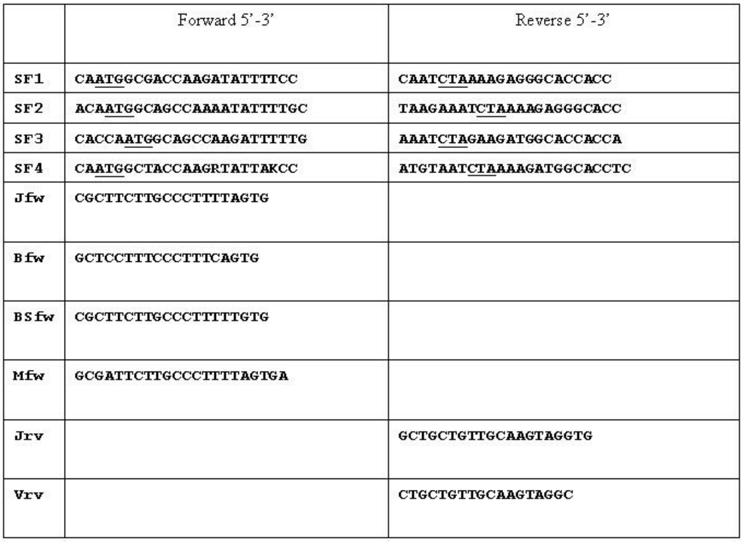
Coding region is in capital letter, 627 nt. Letters in italic identify the coding of the signal peptide: mature polypeptide 188 aa, 21145 Daltons. Stop codon in bold. Double slash indicate the occurrence of the 93 nt deletion in respect to the az19D1 (NM_001111586.1). Underlined bases indicate the five nucleotide changes, in bold those that generate the three aa substitutions. The aa sequence of the mature D3 polypeptide has a 99% identity to the GeneBank accession AI677029.1 cDNA clone of the Ohio43 inbred.

acatagtgaagtacatcagcaacatcttagcacca*ATGGGAGCCAAGATTTTTGCCCTCCTTGCCCTCCTTGCTCTTTCAGCAAGCG*
*CTGCTACCTC*GACTTTTATTCCACAATGCTCACAACAATACCTCTCTCCGG**C**GACAGCCG**T**GGGATTTCAATACCCAACTATACAAT
CCTACATGGTACAAGAGGCCATCCAAGCAAGCATCTTACGGTCATTAGCATTAACCCTCCAACAACCATATGCTCTAT**T**GCAACAGC
CATCCTTAGTGCATCTGTATCTCCAAAGAATCGCGGCACAACAACTACAACAACAGTTGCTACCAACAATCAATCAA//GCACAACA
GCTACTACCATTTAACCAACTTGTCGGGAGCCCTTATGCCTTCTTACTGCAACAACAGCTTCT**A**CCATTCCATCTGCAAGCTGTGGC
AAACATTGCTGCTTTCTTGAGACAACAACATTTGTTGCCATTTTACCCACAGGTTGTGGGAAACATTAATGCCTTCTTGCAACAGCA
ACAATTGCTACCATTCTACCCACAGAATGTGGCAAACATTGTTGCCTTCTTACAACAACAACAATTGTTGCCATTTAGCCAACATGC
TTTGACGAATCCTACCACCTTATTGCAACCGCCCACCATTGGTGGTGCCATCTTC**tag**atttttatgatttatactgtaataataa
agttctcatgctgatatgtgcgacctctcagtaataaagtattagagatctatattttaaaaaaaaaaaaaaaa

S_figure 2. Two-dimensional (2D) fractionation of storage protein polypeptides. Fractionation was carried out as describe in the legend of Fig. 4.



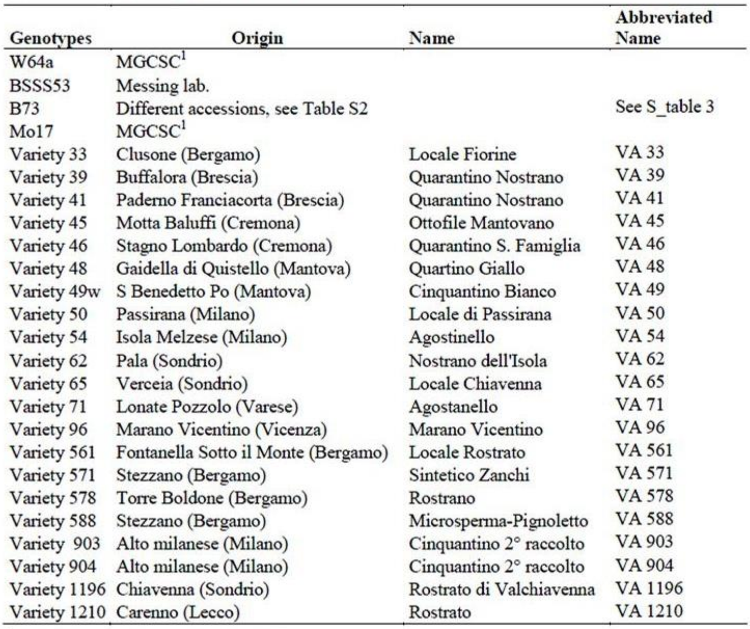
S_Table 1. Primers used in the amplification of alpha-zein subfamilies and analysis of B73 accessions



	Forward 5'-3'	Reverse 5'-3'
SF1	CAATGGCGACCAAGATATTTTCC	CAATCTAAAAGAGGGCACCACC
SF2	ACAATGGCAGCCAAAATATTTTGC	TAAGAAATCTAAAAGAGGGCACC
SF3	CACCAATGGCAGCCAAGATTTTTG	AAATCTAGAAGATGGCACCACCA
SF4	CAATGGCTACCAAGRTATTAKCC	ATGTAATCTAAAAGATGGCACCTC
Jfw	CGCTTCTTGCCCTTTTAGTG	
Bfw	GCTCCTTTCCCTTTCAGTG	
BSfw	CGCTTCTTGCCCTTTTTGTG	
Mfw	GCGATTCTTGCCCTTTTAGTGA	
Jrv		GCTGCTGTTGCAAGTAGGTG
Vrv		CTGCTGTTGCAAGTAGGC

In the SF primers the start codons and the reverse complementary stop codons are underlined.

S_table 2. Maize genotypes (inbreds and varieties)



Genotypes	Origin	Name	Abbreviated Name
W64a	MGCSC[1]		
BSSS53	Messing lab.		
B73	Different accessions, see Table S2		See S_table 3
Mo17	MGCSC[1]		
Variety 33	Clusone (Bergamo)	Locale Fiorine	VA 33
Variety 39	Buffalora (Brescia)	Quarantino Nostrano	VA 39
Variety 41	Paderno Franciacorta (Brescia)	Quarantino Nostrano	VA 41
Variety 45	Motta Baluffi (Cremona)	Ottofile Mantovano	VA 45
Variety 46	Stagno Lombardo (Cremona)	Quarantino S. Famiglia	VA 46
Variety 48	Gaidella di Quistello (Mantova)	Quartino Giallo	VA 48
Variety 49w	S Benedetto Po (Mantova)	Cinquantino Bianco	VA 49
Variety 50	Passirana (Milano)	Locale di Passirana	VA 50
Variety 54	Isola Melzese (Milano)	Agostinello	VA 54
Variety 62	Pala (Sondrio)	Nostrano dell'Isola	VA 62
Variety 65	Verceia (Sondrio)	Locale Chiavenna	VA 65
Variety 71	Lonate Pozzolo (Varese)	Agostanello	VA 71
Variety 96	Marano Vicentino (Vicenza)	Marano Vicentino	VA 96
Variety 561	Fontanella Sotto il Monte (Bergamo)	Locale Rostrato	VA 561
Variety 571	Stezzano (Bergamo)	Sintetico Zanchi	VA 571
Variety 578	Torre Boldone (Bergamo)	Rostrano	VA 578
Variety 588	Stezzano (Bergamo)	Microsperma-Pignoletto	VA 588
Variety 903	Alto milanese (Milano)	Cinquantino 2° raccolto	VA 903
Variety 904	Alto milanese (Milano)	Cinquantino 2° raccolto	VA 904
Variety 1196	Chiavenna (Sondrio)	Rostrato di Valchiavenna	VA 1196
Variety 1210	Carenno (Lecco)	Rostrato	VA 1210

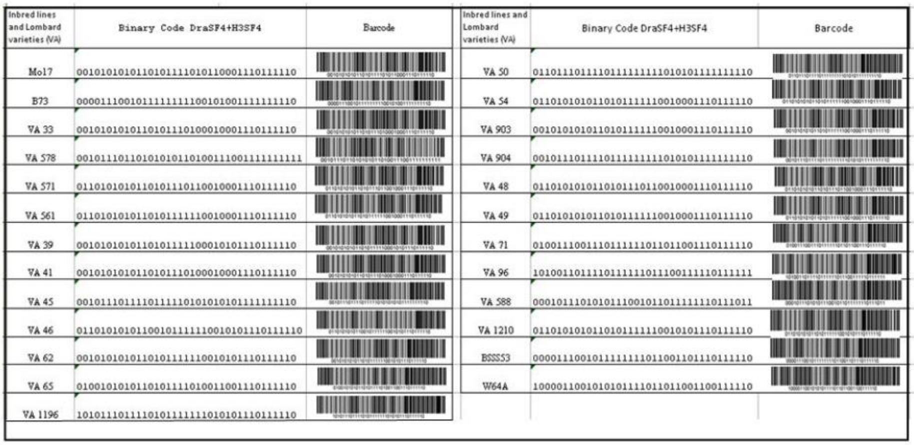
[1] Maize Genetics Cooperation Stock Center

S_table 3 Accessions of B73 inbreds

Year harvest	Origin	Abbreviated Name
1999	Ist. Biologia Biotechnologia agraria	IBB
2008	Centro Ricerche Cerealicoltura, Bergamo	BG
2005	MGCSC[1], C736G	M1
1985	MGCSC[1], 3409-95	M2
2006	MGCSC[1], C736G	M3
1993	Ist. Genetica, Milan University	IG93
2009	Ist. Genetica, Milan University	IG09
2002	Dip. Scienze Agrarie, Bologna University	SAB

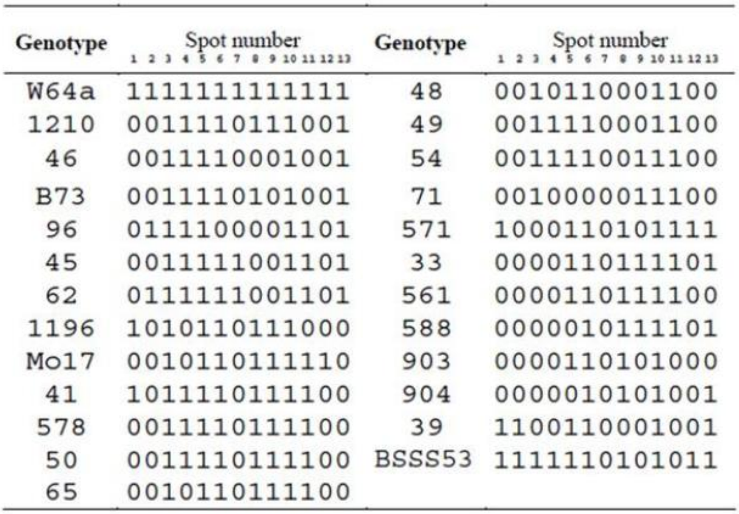
[1]Maize Genetics Cooperation Stock Center

S_table 4 Binary code and barcode of bands of the DNA gel blots of Fig. 2A.



Inbred lines and Lombard varieties (VA)	Binary Code DraSF4+H3SF4	Barcode	Inbred lines and Lombard varieties (VA)	Binary Code DraSF4+H3SF4	Barcode
Mo17	00101010101101011110101100011110111110		VA 50	01101110111101111111110101011111111110	
B73	00001110010111111111100101001111111110		VA 54	01101010101101011111100100011110111110	
VA 33	00101010101101011101000100011110111110		VA 903	00101010101101011111100100011110111110	
VA 578	001011101101010101101001110011111111111		VA 904	00101110111101111111110101011111111110	
VA 571	01101010101101011101100100011110111110		VA 48	01101010101101011101100100011110111110	
VA 561	01101010101101011111100100011110111110		VA 49	01101010101101011111100100011110111110	
VA 39	0010101010110101111100010101110111110		VA 71	01001110011101111111011011001110111110	
VA 41	00101010101101011101000100011110111110		VA 96	10100110111101111111011100111110111111	
VA 45	00101110111101111101010101011111111110		VA 588	00010111010101011100101101111111110111011	
VA 46	011010101011001011111100101011110111110		VA 1210	01101010101101011111100101011110111110	
VA 62	00101010101101011111100101011110111110		BSSS53	00001110010111111111011001101110111110	
VA 65	01001010101101011110100110011110111110		W64A	10000110010101011110110110011001111110	
VA 1196	10101110111101011111110101011110111110				

S_table 5. Binary code of spots from 2D gels reported in Figures 4 and S2



Genotype	Spot number	Genotype	Spot number
	1 2 3 4 5 6 7 8 9 10 11 12 13		1 2 3 4 5 6 7 8 9 10 11 12 13
W64a	1111111111111	48	0010110001100
1210	0011110111001	49	0011110001100
46	0011110001001	54	0011110011100
B73	0011110101001	71	0010000011100
96	0111100001101	571	1000110101111
45	0011111001101	33	0000110111101
62	0111111001101	561	0000110111100
1196	1010110111000	588	0000010111101
Mo17	0010110111110	903	0000110101000
41	1011110111100	904	0000010101001
578	0011110111100	39	1100110001001
50	0011110111100	BSSS53	1111110101011
65	0010110111100		