UNIVERSITÀ DEGLI STUDI DI MILANO

FACOLTÀ DI SCIENZE MATEMATICHE FISICHE E NATURALI

DIPARTIMENTO DI SCIENZE BIOMOLECOLARI E BIOTECNOLOGIE

CORSO DI DOTTORATO DI RICERCA IN SCIENZE GENETICHE E BIOMOLECOLARI (XVII CICLO)

TESI DI DOTTORATO DI RICERCA

ADDRESSING THE MOLECULAR BASES OF HETEROSIS IN MAIZE (*Zea mays* L.)

Dottorando: Giorgio Pea

matr. R04553

Tutor: Chiar.mo prof. M. ENRICO PÈ

Coordinatore: Chiar.mo prof. PAOLO PLEVANI

ANNO ACCADEMICO 2004-2005

Ai miei genitori

Summary

General introduction and project overview

Heterosis in maize

Pioneer studies on hybrids

Hybrid vigor was first recognized as it applies to crop production at the beginning of the last century by Shull (Shull 1908), who coined the term "heterosis" to define the superiority of an F_1 hybrid over its inbred parents. However, genetics of hybrid had actually begun at the end of the previous century, when Charles Darwin, in one of the experiments to test his theory on the origin of species, made a comparison between inbred and cross-pollinated maize. He noticed that the progeny of cross-pollinated maize plants was 25% taller than the progeny of self-pollinated plants and had greater tolerance to cooler growing conditions. From his experiments, he concluded in 1876 that, as a general rule, cross-pollinated (hybrid) plants have "greater height, weight, and fertility" as compared with their self-pollinated counterparts because of their "greater innate constitutional vigor" (Darwin 1876). Later on, in the United States, William Beal at Michigan State College extended Darwin's observations on hybrid vigor, by crossing pairs of open-pollinated varieties of maize. Again he observed increased vigor and grain yield, especially in the hybrids resulting from the crossing of different varieties, and in 1880 he publicly encouraged the systematic use of this method for increased hybrid production (Wallace and Brown 1988). However, the open-pollination method proposed by Beal often gave unpredictable results, and thus it was never exploited on a large scale production.

From theory to practice: towards large scale maize hybrids production

Until the first three decades of the 20th century, improvements of agriculture production were primarily directed towards an increase in cultivated area, the development of the cultivation techniques (machineries and fertilization procedures) and the control of weeds and parasites. However, even the relevant progresses achieved couldn't satisfy the ever increasing market demands for forages and for food quality and quantity. Moreover it became clearer that somehow the improvement of the "environmental conditions" were limited by the "intrinsic properties" within the biological systems themselves. In this contest, the Mendelian principles, bearing to a renewed life, opened the way towards

modern genetics, which were going to provide the producers with new possibilities to answer the socio-economical issues related to the ever increasing demand for food: ever since genetic approaches have played a major role in crop improvement programs, and major realizations in agriculture were achieved by applying breeding procedures, which became more and more sophisticated and effective as genetic tools and knowledge developed. Among such developments, the exploitation of hybrids was one of the most important and effective (Dobzhansky 1950; Bourlaug 2000).

At the turn of the 20th century, urban populations encountered an enormous burst, with a concomitant increase in the demand for meat, which in turn increased demand for feed grains. New cultivable lands were no longer available, and therefore it became clear that an increase in production had inevitably to come from higher yields. The use of breeding methodologies to produce new and/or improved, higher-yielding varieties of maize was at first considered to be a promising option. However, the use of improved varieties, and not hybrids, did not produced the desired effects: the average maize yields in the mid-western Corn Belt state of Iowa, for example, were essentially unchanged during the first three decades of the century (USDA/NASS 1997). In the meanwhile, new bases for the efficient production of hybrids came from the work of Shull and East, who, independently, in 1908 reported their results on the phenomenon of inbreeding depression and hybrid vigor in maize (East 1908; Shull 1908). It was in this contest that Shull coined the term "*heterosis*" to describe hybrid vigor. East and Shull went slightly (but significantly) further than Beal, designing a "controlled" crossing scheme, consisting in several generations of self-pollinations to produce essentially homozygous (pure-breeding) inbred lines that were then crossed to generate hybrids. This new approach had a crucial advantage on the open-pollination method: crosses leading to high-yielding maize hybrids, once identified, could be reproduced without change year after year. Moreover, as opposed to labor-intensive hand pollination, this method was easily manageable: hybrid seeds could be made on a farm-field scale simply removing the tassels from one inbred (detasselling) and allowing it to be pollinated by a second inbred planted in adjacent blocks. In this respect, maize is unique among the cereal crops in that male and female flowers are borne on separate organs (tassel and ear, respectively) and it is wind-pollinated: no other crop is so well suited by nature to large-scale hybrid seed production. However the production process as it was (*two-way* crossing) had a weak point that both the authors themselves recognized: the level of inbreeding depression in the inbred lines had so dramatic effects on the plants that production of hybrid seeds was seriously compromised, resulting in a

significant cost increase. In other words, the better performance of hybrids did not cover the cost for purchasing elite hybrid (Duvick 2001).

Despite these not promising results, the inbred-hybrid idea did not die, and maize was going to become very soon the economically most exploited and studied hybrid culture worldwide. First solutions to a more affordable and effective hybrid seed production came in fact within the next years from alternative crossing designs, as for example the *three-* and *four-way* hybrid production schemes (Figure 1). Even if in such hybrids the heterotic effects were generally more limited than for the *two-way* ones, nevertheless the best hybrid developed by these crossing designs was always superior to the best open-pollinated variety. In simple terms, the general strategy was to balance at best the performance due to heterosis and its reproducibility on one side, while limiting the constraints of inbreeding depression on the other.
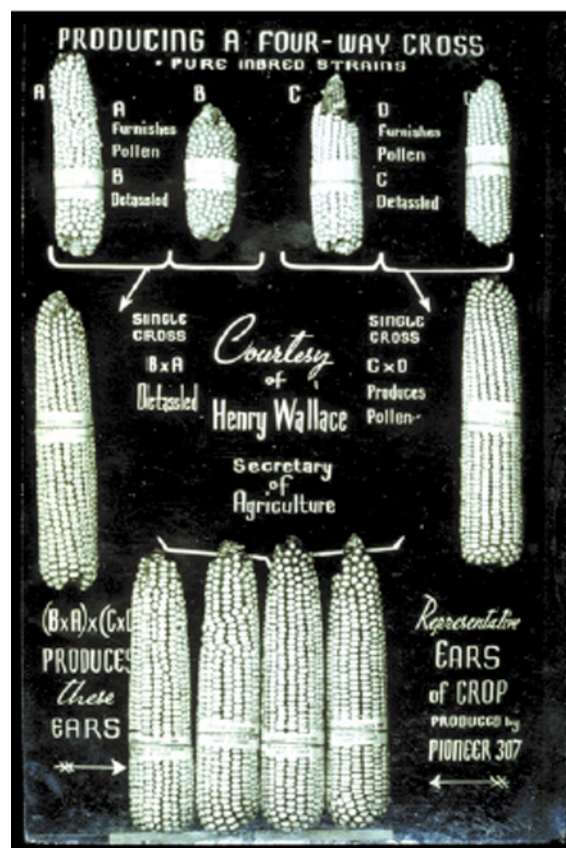

Figure 1: Illustration, from a farm magazine in the 1930s, showing the scheme for the production of four-way maize hybrids (font: Duvick, 2001).

Maize breeders, at least initially from both the public and the private sector, continually turned out with higher-yielding hybrids, year after year, by applying and refining their mating designs, and by 1960 virtually all maize plantings in the United States were hybrids (Figure 2). In 1997 United States maize yields averaged 8ton/h, compared with 1ton/h in 1930 (USDA/NASS - Crop Production Data, Washington, DC, 2000). Of course, the use of synthetic nitrogen fertilizers, chemical weed control products, and more efficient cultivation and harvesting techniques, mostly introduced around the 1950s, have also contributed to this increment in yields (Cardwell 1982; Castelberry et al. 1984; Russel 1991; Duvick 1992).
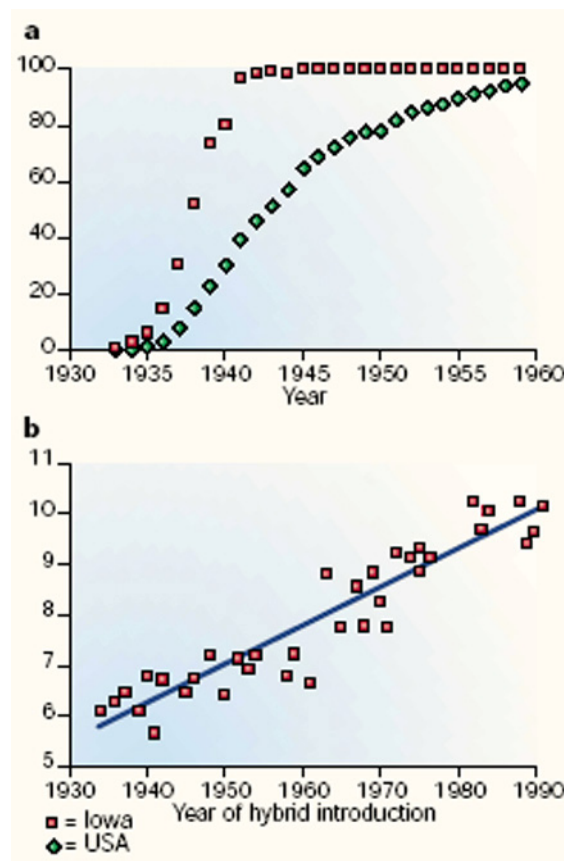
Figure 2: Maize hybrids production. **a**: Percent of maize cultivated area planted to hybrids from 1930 to 1960 in Iowa (red) and in the United States (green); **b**: Grain yields (tons/ha) of different maize hybrids introduced in central Iowa from 1934 to 1991.

Surprisingly, several experiments have shown that the levels of heterosis was actually unchanged over these years, while the yields of the inbred lines rose at almost the same rate as those of the hybrids (Duvick 1999). It seems therefore that yield gains have primarily come from genetic improvements in tolerance to different stresses in the maize

germplasm (disease, insects, plant density, drought, low soil fertility) which have then been combined, stabilized and further enhanced in the hybrids production programs. By the 1960s, the new inbred lines were so improved that it became practical to use them as convenient seed holder for the production of high-performing two-way hybrids (Figure 3). These yielded more than the best three- and four-way hybrids and were now feasible for commercialization due to the lower cost of seed production.

Figure 3: Representative individuals from two elite maize inbred lines, B73 (far left) and Mo17 (far right) and the progeny of their reciprocal hybrid crosses, B73/Mo17 (left center) and Mo17/B73 (right center); the female parent is listed first in maize genetics nomenclature. The hybrid progeny of the cross between these two lines are clearly taller and more productive than either parent, illustrating the concept of heterosis. B73 and Mo17 are high-quality inbred lines, largely employed in the 1970s for the production of single-cross high-performing hybrids.

Big seed companies, capable of developing and producing new varieties, as well as of selling the hybrid seeds to the farmers on a large scale, more and more controlled the business, obviously at the expenses of the smallest ones. The exchange of information and materials among private and public sector that had characterized the pioneering era of maize improvement also changed radically: seed companies kept the pedigrees of their hybrids secret and stopped trading their inbred lines; while public sector, conversely,

progressively shifted from the development of the commercial inbreds to the study of the theoretical basis for producing improved inbreds and hybrids.

Nowadays, high-performing maize commercial hybrids are cultivated in all the production areas of the industrialized Countries such as North America and Europe, as well as in the industrial agriculture regions of Developing Countries such as Argentina, China and Brazil (Duvick 2001).

Genetic and molecular hypotheses for heterosis

Classic genetic hypotheses for heterosis

Shull's original definition of heterosis simply describes the improved phenotype resulting from the crossing between different parental inbred lines, but does not explain its genetic basis (Lamkey and Edwards 1999). However, it was immediately clear that if these had been known, breeding programs could have been more precise and hybrid yields presumably could have been advanced further. Genetic theories were proposed and experiments conducted toward this goal.

Three major classical genetic models have been suggested to explain the hybrid vigor: dominance, real or pseudo overdominance and epistasis. The dominance hypothesis attributes increased vigor to the action at multiple loci of favorable dominant alleles from both parents combined in the hybrid (Bruce 1910; Jones 1917; Xiao et al. 1995; Cockerham and Zeng 1996). The overdominance hypothesis postulates instead the existence of loci at which the heterozygous state is superior to either homozygotes (Shull 1908; East 1936; Crow 1948; Stuber et al. 1992); pseudo-overdominance refers to a particular situation, in which tightly linked genes with favorable dominant alleles in repulsion phase in the parental lines result in an apparent overdominance when combined in the hybrid. Finally, the interaction of favorable alleles from the two parents at different loci, themselves showing additive, dominant or overdominant actions, is taken in account by the epistasis hypothesis (Stuber et al. 1992; Li et al. 2001; Luo et al. 2001).

Genetic approaches to uncover and predict heterosis

Initially, the classical genetic studies employed morphological traits (mutations) as genetic markers. This generally posed major limitations on the power of the analyses, since only few markers could be followed in any given cross and the markers themselves could affect plant traits, thus producing confounding phenotypic effects (Tanksley et al. 1982;

Stuber 1992). In more recent years, the introduction of molecular markers, and particularly those based on the detection of DNA polymorphisms, have provided geneticists with a wealth of phenotypically neutral markers, particularly powerful for studying inheritance of quantitative traits. Basically, dense-mapped genetic markers could be used for identifying quantitative trait loci (QTLs) controlling relevant complex traits, providing the bases for the comprehension of key phenomena that can subsequently be applied to plant improvement programs. These techniques have also been widely applied in the attempt to shed light on hybrid vigor. However, to-date they have not produced data consistent with a unique genetic explanation for the phenomenon of heterosis in maize or in any other species (Coors and Pandey 1999). Moreover, molecular marker approaches have been extensively applied in breeding programs, as a tool for the selection of the best parental lines to be crossed (marker assisted selection); in the specific case of hybrid production, they were employed as potential instruments for *a priori* prediction of the hybrid performance from a given cross. Among other parameters, parental genetic distance has been regarded as a possible indicator for hybrid performance (Melchinger 1999), and the development of molecular marker systems such as AFLPs, SSRs, and SNPs considerably facilitated and improved the power of the genetic distance estimation between genotypes. Several studies have in fact reported a positive correlation between genetic distance and heterosis in maize (Liu et al. 2002; Barbosa et al. 2003). However, studies on other plant species often failed to detect a relationship between these two parameters (Cerna et al. 1997; Joyce et al. 1999; Liu et al. 1999; Riday et al. 2003); further, heterosis in maize has been reported to culminate at an optimum of parental genetic distance before declining again (Moll et al. 1965). As a matter of fact, no prediction parameters have yet been fully related to heterosis and the selection of the lines for highly-performing hybrid production is still based on an empirical evaluation of the performance of the hybrid progeny. It is not unlikely, however, that the design of appropriate genetic schemes, as well as the continuous improvement of the statistical tools for the analysis of QTL, might in future lead to a more precise dissection of complex traits, and thus also to a more accurate and reliable insight into the genetic basis of such a complex phenomenon as heterosis.

Theories on the molecular bases of heterosis

In addition to genetic hypotheses, numerous physiological and molecular mechanisms underlying heterosis were also proposed (Birchler et al. 2003). One of the first explanations formulated for heterosis was that when the hybrid is produced, all the

different slightly deleterious alleles at multiple loci present in the two parental inbred lines are complemented in a dominant way, thus generating a progeny that exceeds either parents. An early criticism to this idea was that, if this hypothesis were true, it should be possible to create an inbred line carrying all of the superior alleles, a situation that has not actually occurred. In favor of the complementation hypothesis it could be argued that, with so many genes involved, this could be due to an extremely low probability of accumulating all of the better alternatives into one line, as an effect of linkage between deleterious alleles and superior alleles of different genes. However, even if any deleterious alleles would become homozygous in different inbred lines, gene complementation in hybrids might account for only the hybrid being equivalent to the better of the two parents. Alternatively, interactions of alleles of the same genes (or from different genes) may be invoked: assuming that their effects were cumulative in the phenotype, then heterosis would result. In fact, several observations suggest that the basic principle of heterosis is something other than simple complementation. The strongest evidence is that although inbred lines have been improved greatly over the decades, the magnitude of heterosis, defined as the difference in yield between a single cross hybrid and the mean of its two inbred parents, has not diminished but has rather been maintained or even slightly increased (East 1936; Duvick 1999); whereas, if hybrid vigor were merely caused by the complementation of deleterious alleles, and if the improved inbred lines have been progressively purged of the most severe of such alleles, then the absolute amount of heterosis might be expected to decline. A further indication against the complementation hypothesis comes from the fact that, as mentioned, the quality of two inbred lines does not necessarily predict the amount of heterosis in their hybrid, which must still be determined in a cross. In fact, all the previous observations rather suggest that the slight increase in hybrid vigor over the years might have occurred by selection of alleles at the right set of loci producing the best combinations to bring about heterosis.

Given that, since quantitative traits are in large part under the control of multiple dosage-dependent regulatory loci, it could be hypothesized that heterosis could result from different alleles being present at loci contributing to the plant regulatory hierarchies. Indeed, recent studies indicate that the expression of many genes in hybrids does not exhibit the expected midparent value (Romagnoli et al. 1990; Leonardi et al. 1991; Osborn et al. 2003; Song and Messing 2003), suggesting that a shift in gene regulation between inbred lines and hybrids could account for hybrid vigor. Remarkably, in all the cited studies, the range of relative deviation in gene expression between the inbred lines and the

hybrids fell within a twofold change, i.e. small differences in the expression of many genes was observed rather than the opposite. However, it is not yet clear if the observed differences in gene expression could be responsible for heterosis, or a result of it. A large survey of gene expression in inbred lines and hybrids, providing some answer about the spectrum of genes that are influenced and the direction the changes occur, might help understanding which genes could be possibly involved and how the heterozygosity at loci could influence heterotic response.

Generally speaking, two extreme mechanisms have been proposed at this level: i) a combined allelic expression of various genes in the hybrid; ii) a regulatory gene allelic interaction of many genes that, in the hybrid, determines the deviation of gene expression from midparent prediction (Birchler et al. 2003). In fact, allelic expression variation has been suggested to play important roles in determining phenotypic diversity, since regulatory allelic variants can affect the level of gene expression and result in quantitative variants. In particular, allelic diversity has been recently proposed to be an important genetic component for phenotypic variation especially in plants (Doebley and Lukens 1998; Buckler and Thornsberry 2002). The huge amount of nucleotide sequence, made available in recent years by the advancements in sequencing technology, has given the possibility of fine scale sequence comparisons: data are revealing that nucleotide sequences variation widely exists not only between, but also within species. Noticeably, maize genome has revealed an extremely high level of DNA sequence polymorphism, which has been estimated an order of magnitude higher than that observed in human (Sunayaev et al. 2000; Bhattramakki et al. 2002; Buckler and Thornsberry 2002; Ching et al. 2002); the differences range from single nucleotide polymorphisms (SNPs) to large regions of several kilobases (Fu and Dooner 2002). These variations in allelic regions may indeed play a role in gene regulation, especially in hybrids, where pairs of differently regulated alleles are coupled and can actively interact. Recently, a relationship between heterosis in maize and differential allelic expression resulting from different regulatory region (cis elements) has been proposed (Guo et al. 2004): 11 out of the 15 analyzed genes showed differences in the levels of allelic transcripts in different maize hybrids, ranging from a perfect biallelic to a monoallelic expression; moreover, in some cases, allelic transcription levels were modulated differently by environmental stresses in differently performing hybrids. These results suggest that the combination of different functionalities in parental alleles within the hybrid might have a major impact on heterosis. Furthermore, in their sequence-based investigation, Fu and Dooner (2002) have reported that not only a

variety of intergenic differences were present between collinear regions of different maize inbred lines, but interestingly, also four of the predicted genes in one of the analyzed haplotypes were missing in the other. Analogous results were obtained in similar surveys conducted on the *c1Z* cluster locus (Song and Messing 2003) and on four additional large collinear regions between B73 and Mo17 inbred lines (Brunner et al. submitted), suggesting that this could be a common feature of the maize genome. If this is the case, not only the different regulation of the parental alleles, but also the presence of non-shared genes, if they are transcriptionally functional, might contribute in creating a unique transcriptome in the hybrids. In other words, these observations might suggest that hybrids could actually inherit a unique gene complement deriving from the combination and the interaction of different parental genomes. As a consequence, these features, which in any case may not be fixed in any homozygous inbred lines, would be peculiar of hybrids and could possibly account for heterosis in maize.

Project overview

Even though heterosis is widely exploited in agriculture for the production of hybrid varieties, high-performing in terms of productivity and quality, its genetic and molecular bases still remain basically unknown. However, the identification of the molecular mechanisms and of the genetic interactions responsible for hybrid vigor might allow the development of new molecular tools both for the evaluation of germplasm to be employed in crop breeding and to increase selection efficiency.

In this thesis I report the results of experiments carried out with the purpose of contributing to a deeper understanding of the molecular mechanism underlying heterosis in maize, by investigating the possible role of the modulation of genic expression in determining the hybrid vigor. My data are reported in the form of full paper manuscript already submitted or in the process of being submitted.

The first manuscript deals with a genome-wide comparison of gene expression in immature ears carried out using cDNA microarray technology, in the attempt to reveal the relationships between differences in gene expression levels between inbred lines and their hybrid and heterosis, as well as to detect gene functions potentially involved in it.

The second manuscript reports my experimental activity on the expression analysis of intra-specific non-shared genes from large allelic genome segments between two different maize inbred lines (B73 and Mo17), which was conducted during a stage period in the laboratory of Antoni Rafalski at DuPont Experimental Station (Wilmington, DE, USA).

Finally, I actively collaborated in the analysis of allele specific expression in maize hybrids, whose results are reported in the third manuscript. In particular, the relative amount of parental-specific allele transcripts was monitored in different tissues and different condition within the same hybrid, in order to uncover possible regulatory properties that could underlie heterosis.

References

Barbosa, A.M.M., Geraldi, I.O., Garcia, A.A.F., Souza, C.L. and Souza, A.P. (2003). "Relationship of intra- and interpopulation tropical maize single cross hybrid performance and genetic distances computed from AFLP and SSR markers." Euphytica **130**: 87-89.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J.C., Tingey, S. and Rafalski, A. (2002). "Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers." Plant Mol Biol **48**: 539-547.

Birchler, J.A., Auger, D.L. and Riddle, N.C. (2003). "In search of the molecular basis of Heterosis." The Plant Cell **15**: 2236-2239.

Bourlaug, N.E. (2000). "The Green Revolution revisited and Road Ahead." Presentation for the 30th anniversary of the Nobel Prize for Literature.

Bruce, A.B. (1910). "The Mendelian theory of heredity and the argumentation of vigor." Science **32**: 627-628.

Buckler, E.S. and Thornsberry, J.M. (2002). "Plant molecular diversity and applications to genomics." Curr Opin Plant Biol **5**: 107-111.

Cardwell, V.B. (1982). "Fifty years of Minnesota corn production: sources of yield increase." Agronomy J. **74**: 984-990.

Castelberry, R.M., Crum, C.W. and Krull, C.F. (1984). "Genetic yield improvement of U.S. maize cultivars under varying fertility and climatic environments." Crop Sci. **24**: 33-36.

Cerna, F.J., Cianzio, S.R., Rafalski, A., Tingey, S. and Dyer, D. (1997). "Relationship between seed yield heterosis and molecular heterozygosity in soybean." Theor Appl Genet **95**: 460-467.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M. and Rafalski, A. (2002). "SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines." BMC Genet **3**: 19.

Cockerham, C.C. and Zeng, Z.B. (1996). "Design III with marker loci." Genetics **143**: 1437-1456.

Coors, J.G. and Pandey, S. (1999). Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA.

Crow, J.F. (1948). "Alternative hypotheses of hybrid vigor." Genetics **33**: 477-487.

Darwin, C. (1876). The effects of cross and self fertilization in the vegetable kingdom. New York Univ. Press, 1989.

Dobzhansky, T. (1950). "Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of Drosophila pseudobscura." Genetics **35**: 288-302.

Doebley, J. and Lukens, L. (1998). "Transcriptional regulators and the evolution of plant form." Plant Cell **10**: 1075-1082.

Duvick, D.N. (1992). "Genetic contributions to advances in yield of U.S. maize." Maydica **37**: 69-79.

Duvick, D.N. (1999). In: Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA**:** 19-29.

Duvick, D.N. (2001). "Biotechnology in the 1930s: the development of hybrid maize." Nat Rev Genet **2**(1): 69-74.

East, E.M. (1908). "Inbreeding in corn." Rep Connecticut Agric Exp Stn for 1907**:** 419-428.

East, E.M. (1936). "Heterosis." Genetics **21**: 375-397.

Fu, H. and Dooner, H.K. (2002). "Intraspecific violation of genetic colinearity and its implications in maize." PNAS **99**(14): 9573-9578.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A. and Smith, O.S. (2004). "Allelic Variation of Gene Expression in Maize Hybrids." Plant Cell **16**(7): 1707-1716.

Jones, D.F. (1917). "Dominance of linked factors as a means of accounting for heterosis." Genetics **2**: 466-479.

Joyce, T.A., Abberton, M.T., Michaelson-Yeates, T.P.T. and Forster, J.W. (1999). "Relationships between genetic distance measured by RAPD-PCR and heterosis in inbred lines of white clover (Trifolium repens L.)." Euphytica **107**: 159-165.

Lamkey, K.R. and Edwards, J.W. (1999). The quantitative genetics of heterosis. In: Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA**:** 31-48.

Leonardi, A., Damerval, C., Herbert, Y., Gallais, A. and DeVienne, D. (1991). "Association of protein amount polymorphisms (PAP) among maize lines with performances of their hybrids." Theor Appl Genet **82**: 552-560.

Li, Z.-K., Luo, L.J., Mei, H.W., Wang, D.L., Shu, Q.Y., Tabien, R., Zhong, D.B., Ying, C.S., Stansel, J.W., Khush, G.S. and Paterson, A.H. (2001). "Overdominant Epistatic Loci Are the Primary Genetic Basis of Inbreeding Depression and Heterosis in Rice. I. Biomass and Grain Yield." Genetics **158**(4): 1737-1753.

Liu, X.C., Ishiki, K. and Wang, W.X. (2002). "Identification of AFLP markers favorable to heterosis in hybrid rice." Breed Sci **52**: 201-206.

Liu, Z.Q., Pei, Y. and Pu, Z.J. (1999). "Relationship between hybrid performance and genetic diversity based on RAPD markers in wheat, Triticum aestivum L." Plant Breed **118**: 119-123.

Luo, L.J., Li, Z.K., Mei, H.W., Shu, Q.Y., Tabien, R. and Zhong, D.B. (2001). "Overdominant Epistatic Loci Are the Primary Genetic Basis of Inbreeding Depression and Heterosis in Rice. II. Grain Yield Components." Genetics **158**(4): 1755-1771.

Melchinger, A.E. (1999). Genetic diversity and heterosis. In: Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA**:** pp 99-118.

Moll, R.H., Lonnquist, J.H., Fortuna, J.V. and Johnson, C.E. (1965). "The relationship of heterosis and genetic divergence in maize." Genetics **52**: 139-144.

Osborn, T.C., Pires, J.C., Birchler, J.A., Auger, D.L., Chen, Z.J., Lee, H.S., Comai, L., Madlung, A., Doerge, R.W., Colot, V. and Martienssen, R.A. (2003). "Understanding mechanisms of novel gene expression in polyploids." Trends Genet **19**: 141-147.

Riday, H., Brummer, E.C., Campbell, T.A., Luth, D. and Cazcarro, P.M. (2003). "Comparisons of genetic and morphological distance with heterosis between Medicago sativa subsp sativa and subsp falcata." Euphytica **131**: 37-45.

Romagnoli, S., Maddaloni, M., Livini, C. and Motto, M. (1990). "Relationship between gene expression and hybrid vigor in primary root tips of young maize (Zea mays L.) plantlets." Theor Appl Genet **80**: 769-775.

Russel, W.A. (1991). "Genetic improvement of maize yields." Advances in Agronomy **46**: 245-298.

Shull, G.H. (1908). "The composition of a field of maize." Rep Am Breeders Assoc **4**: 296-301.

Song, R. and Messing, J. (2003). "Gene expression of a gene family in maize based on non-collinear haplotypes." Proc Natl Acad Sci USA **100**: 9055-9060.

Stuber, C.W. (1992). "Biochemical and molecular markers in plant breeding." Plant Breed. Rev. **9**: 37-61.

Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T. and Lander, E.S. (1992). "Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers." Genetics **132**: 823–839.

Sunayaev, S.R., Lathe, W.C., Ramensky, V.E. and Bork, P. (2000). "SNP frequencies in human genes: An excess of rare alleles and differing modes of selection." Trends Genet **16**: 335-337.

Tanksley, S.D., Medina-Filho, H. and Rick, C.M. (1982). "Use of naturally-occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato." Heredity **49**: 11-25.

USDA/NASS (1997). Iowa Agricultural Statistics (NASS/USDA, Washington DC).

Wallace, H.A. and Brown, W.L. (1988). Corn and Its Early Fathers. Ames, IA, USA, Iowa State Univ. Press.

Xiao, J., Li, J., Yuan, L. and Tanksley, S.D. (1995). "Dominance is the major genetic basis of heterosis in rice as revealed by QTL analysis using molecular markers." Genetics **140**: 745-754.

Transcriptional profiling of immature ears in a maize F_1 hybrid and in its corresponding inbred lines

Giorgio Pea, Simona Ferron, Luca Gianfranceschi, and M. Enrico Pè

Department of Biomolecular Sciences and Biotechnology, University of Milan, 20133 Milan, Italy

Abstract

The genetic and molecular mechanisms underlying heterosis are still unclear. However, recent data suggest that regulation of gene expression could play an important role in determining hybrid vigor. As a contribution to uncover genes and mechanisms possibly causing or being influenced by heterosis, here we present data on the transcription profiling in immature ears between inbred lines B73 and H99 and their corresponding F_1 hybrid using cDNA microarray technology and Real Time PCR. The relative expression of 4905 ESTs represented in triplicate, corresponding to about 1700 maize genes, was investigated simultaneously on five replicates (for a total of 15 data points for each EST in each comparison). Relative variation of gene expression generally did not exceed a ±1.5-fold value. However, using two different statistical approaches, we were able to identify genes expressed at a significantly different level between both inbred lines and their hybrid; 95% of the called out genes were confirmed beyond a +/-2SD threshold, assuming an overall normal distribution of the normalized ratios across replicates. Both up and down-regulated genes in the hybrid were found, B73 vs. F_1 comparison showing a higher number of differentially expressed genes than the H99 vs. F_1 one. A few ESTs shows the same direction of regulation in both comparisons, suggesting that they could be inbred/hybrid specifically regulated. The absolute expression levels of 3 ESTs for each category of expression for each comparison were also determined by real-time PCR, and 10 out of the 18 tested ESTs confirmed the microarray hybridization data. A putative function was also assigned to the regulated ESTs. Here we discuss the possibility that the observed changes in gene expression between parental lines and their hybrid might be correlated with the heterotic phenomenon.

Introduction

The term heterosis describes the superiority of an F_1 hybrid over its parents. The increased productivity that results from heterosis, combined with the expression of adaptive traits such as increased fertility and resistance to biotic and abiotic stress (Dobzhansky 1950), is exploited through the development of hybrid varieties in several crop species, most markedly in maize (Duvick 2001). Heterosis as it applies to crop breeding was first recognized by Shull in 1908 (Shull 1908). Two major genetic hypotheses for heterosis have been proposed since, the dominance hypothesis (Davenport 1908) and the over-dominance hypothesis (East 1908; Shull 1908). In addition to these, several physiological and molecular processes underlying heterosis have been investigated (Comings and MacMurray 2000; DeVienne et al. 2001); however, due to the difficulties in producing reliable associations between phenotypic effects and multiple-interacting molecular events likely to occur in hybrid vigor, little is understood regarding the molecular basis of the phenomenon and the substantial essence of hybrid vigor is still elusive (Coors and Pandey 1999).

Genetic distance between parents has been proposed as a useful indicator for hybrid vigor prediction (Melchinger 1999). Several studies reported a positive correlation between genetic distance of parental lines and superior hybrid performance (Liu et al. 2002; Barbosa et al. 2003). However, at least in maize, heterosis seems to decline beyond a variable optimum of parental genetic distance (Moll et al. 1965) and new hybrid production still relies basically on cross-and-select empirical, time-consuming approaches. Therefore any added insight that could lead to reliable tools for hybrid performance prediction would have an enormous impact.

Since a relevant part of biological regulations occurs at the transcriptional level, it is not unreasonable to postulate that gene expression, together with other possible regulations, might substantially influence heterosis.

Two extreme models could explain heterosis in terms of gene expression: i) when two different alleles of various genes are joined in the hybrid, there is a combined allelic expression that results in qualitative relevant differences. Recently, unusually high level of allelic transcription variation due to cis-regulatory elements have been independently reported in maize by Guo and coworkers (Guo et al. 2004) and Morgante and coworkers (personal communication), suggesting a possible role of allelic transcription regulation in hybrid vigor; ii) the combination of different alleles at regulatory loci produces interactions

leading to a deviation in gene expression levels in the hybrid from the mid-parent prediction. In a study on the zein *zC1* locus, it was found that of the 10 genes analyzed, only in one case hybrid expression followed the level predicted by the calculated allelic dosage (Song and Messing 2003). In none of similar studies conducted on maize big changes in expression levels were found between inbred lines and their hybrids (Romagnoli et al. 1990; Leonardi et al. 1991; Osborn et al. 2003). This is expected if multiple regulatory loci are involved, which are normally strictly regulated and do not greatly alter their basic expression levels. Consequently, there are reasonable indications for assuming that slight differences could play a major biological role in complex traits variation. In other words, applying quantitative genetics concepts to the molecular level, little effects of numerous figurants, rather than oversize effects of a few main characters, should realistically be predicted to be responsible for the genetic and molecular components of such a complex phenomenon as heterosis. However, despite quantitative genetics methods succeed in partitioning environmental and genetic effects into variance components, their results are often difficult be integrated with developmental/physiological events. In this context, precise quantification of intracellular processes such as transcription should lead to the important goal of joining quantitative genetics to genomic analysis (Kerr and Churchill 2001; Wolfinger et al. 2001), i.e. by monitoring gene expression changes on a large scale, it might be possible to gain crucial information on some of the molecular events underlying heterosis. However, so far, no high-throughput genome-wide surveys have been performed to this end (Birchler et al. 2003).

With the objective of contributing to a molecular rationalization of heterosis in maize, we applied cDNA microarray technology and real-time PCR approaches to compare transcriptional levels between inbred lines B73 and H99, and their corresponding F_1 hybrid, which shows high level of heterosis. In particular we focused on gene expression in ear, as the organ directly involved in yield potential in maize.

Results

Experimental design

Recent studies suggested that different expression of functional associated genes might contribute to heterosis in maize (Song and Messing 2003). To investigate the presence of differences in transcriptional levels between inbred lines B73 and H99 and their heterotic hybrid, we employed cDNA microarray technology. To this end, we directly

contrasted both B73 and H99 inbred lines vs. F_1 hybrid relative gene expression in immature ear. Differences in transcriptional levels between B73 and H99 were inferred using the hybrid as a reference sample in an indirect experimental design (Yang and Speed 2002). In order to minimize variability in transcript population among individuals, total RNA coming from different isolations, each collected on multiple individuals, were mixed before poly(A^+) RNA purification. All hybridizations on microarray slides were then performed using cDNA independently labeled from the poly(A^+) RNA purification product for each genotype. Our experimental design consisted of 10 cDNA microarrays hybridizations, 5 for each combination of hybrid vs. inbred genotypes, involving 20 separate labeling reactions. Since all the ESTs are spotted in triplicate on each slide and 5 independent replicates for each comparison were performed, each data point submitted to subsequent analyses derived from 15 records. The Cy3™ and Cy5™ dyes were also swapped in 2 of the 5 replicates for each combination. Control channel was always assigned to the hybrid.

We determined the Tentative Contig to which each of the spotted EST belongs by applying a query algorithm to the list of maize tentative contigs (TC) available at the TIGR Maize TC annotator (TIGR Maize Gene Index, release 13.0; http://www.tigr.org/tdb/tgi/zmgi/). The analysis revealed that the 4,905 ESTs spotted on the microarrays represent about 2,200 different TCs, among which about 87% resulted homologue to known sequences in database. The homology data allowed estimating that about 1,700 putative different genes are represented in the 606 microarray slides (Table 1).

Total spotted cDNA	5093
ESTs (spotted in triplicate)	4905
Controls (blanks excluded)	188
ESTs belonging to a TC*	4517
w/ known homology	4172
w/o homology	345
singletons	311
ESTs retired from GenBank	77
Total different TCs*	2216
w/ known homology	1931
w/o homology	285
Putative different genes**	1697

* based on TIGR db; ** based on homology

Table 1: Summary of the features of the ESTs spotted on used microarray slides. Each EST present on the slides was assigned to its relative tentative contigs (TCs) querying the ZM Gene Index database available at the TIGR website (http://www.tigr.org, release 13.0). The number of putative different genes has been estimated on the homology data reported for each TC within the same database.

Data normalization

Normalization steps were performed to standardize microarray data and to allow discrimination between biological variations in gene expression levels and experimental errors. Data for each slide were also scaled so that relative gene expression levels could be compared. Raw data from images quantification for each slide, previously purged from bad spots data by manual editing of QuantArray output files, were entered in GeneSpring software. Genes reported multiple times with the same name on different horizontal lines in data file are automatically considered by GeneSpring as replicate measurement and their relative signals are averaged. This software then records the average value for each data, keeping track of the minimum and maximum values, and assuming that the entered data are raw and must be normalized. Data were normalized both at the gene and at the chip level, in order to standardize the expression levels between genes and at between arrays respectively. The application of the normalizing procedure to the data succeeded in centering the median of the ratio values of each array around 1, providing protection from both preparation and incorporation artifacts (Figure 1). Normalization between different slides also allowed performing a statistical correlation analysis between experiments. For both H99 vs. F_1 and B73 v. F_1 comparisons good data reproducibility among replicates was observed (data not shown).
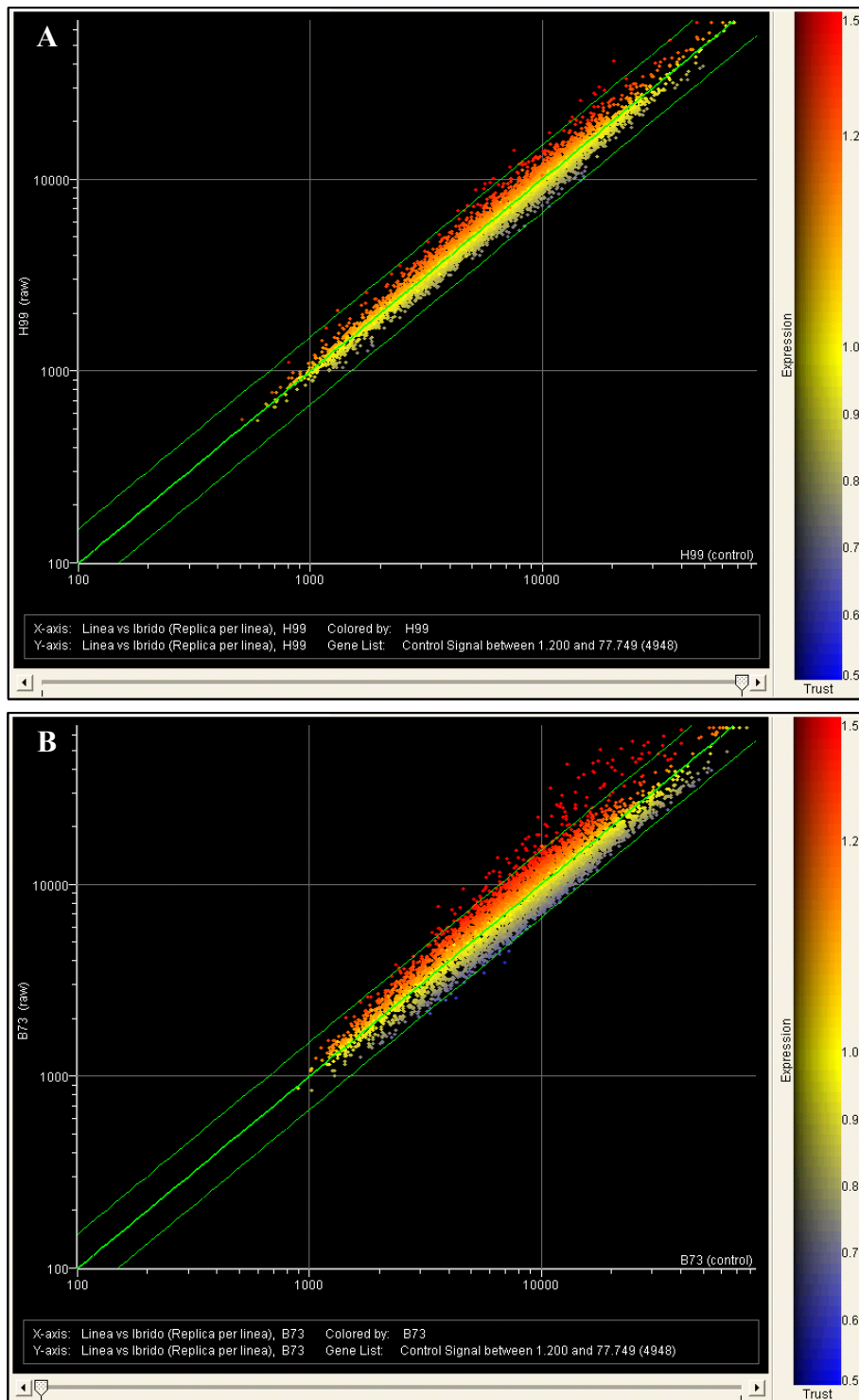
Figure 1: Scatter plot of the normalized fluorescence intensities across replicates (details in the text). Each spot represent the ratio between the sample (parental line) and the control (hybrid) intensity. The central diagonal line indicates the points for which the expression is equal within the compared samples (y=x); the upper and lower lines represent the twofold change cutoffs. Different ratio values are visualized in color scale, with reference to the color bar on the right of each graph; spot brightness indicates the level of reproducibility of each data across replicates (Trust), as determined by statistical correlation analysis. **A:** H99 vs. F_1; **B:** B73 vs. F_1. X-axis: control channel intensity (F_1); Y-axis: sample channel intensities (parental line); axes are in logarithmic scale.

Statistical analysis

For statistical analysis, the normalized ratio values relative to each of the 5 hybridizations were exported from GeneSpring and submitted to SAM software (Tusher et al. 2001). This software relies on a statistical approach for finding significant genes in a set of microarray experiments: performing repeated permutations of the data, it computes a statistic measuring the strength of the relationship between gene expression and the response variable to determine if the expression of any of the genes is significantly related to the response. The significance cut-off may be selected by the user according to the accepted false positive rate. The number of replicate data points for each EST (15) allowed us to set the level of significance according to stringent statistical criteria, i.e. imposing median of false significant rate value less or equal to one.

None of the EST normalized ratios showed values that exceeded ±1.5 fold-change (Figure 1). However, sets of ESTs for each comparison were called out by the SAM software as significantly differentially expressed (Figure 2), with a ratio cut-off value corresponding approximately to a ±1.15 fold change. Up and down-regulation are always referred as that of the inbred lines compared to the hybrid. A higher number of differentially expressed ESTs were called out in B73 vs. F_1 comparison: 1,160 ESTs (764 up-regulated and 396 down-regulated, Figure 2A), versus a total of 191 ESTs (130 up-regulated and 61 down-regulated) detected in H99 vs. F_1 (Figure 2B).
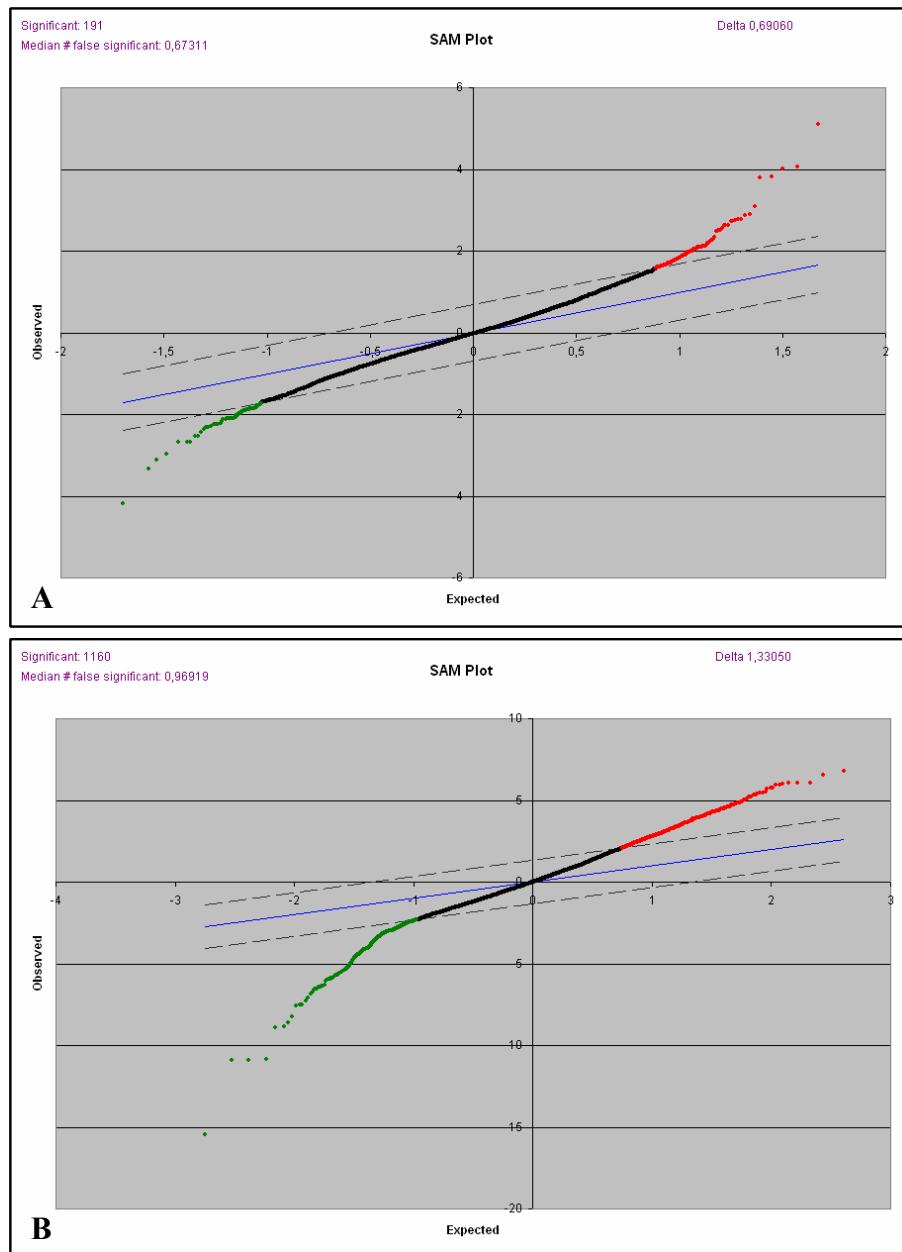
Figure 2: Graphical output of SAM statistical analysis of microarray data (details in the text). Significant up and down-regulated genes are displayed as red and green spots respectively. **A:** H99 vs. F_1; **B:** B73 vs. F_1.

A confidence-filtering analysis, based on the Bonferroni and Hockberg corrected t-test for multiple samples, was also applied on the microarray dataset, producing a second list of significantly regulated ESTs for each comparison. The relative t-test significance cut-off p-values for H99 vs. F_1 and B73 vs. F_1 corresponded to 15.4% and 4.1% respectively. When compared, 78.5% and 90.8% of ESTs were found in common between the two statistical analyses within H99 vs. F_1 and B73 vs. F_1 comparison respectively (Figure 3A). SAM software showed a propensity for calling out a higher number of ESTs down-regulated in parental lines (H99<F_1 and B73<F_1), while an opposite trend was

observed for ESTs up-regulated in parental lines (Figure 3B). A conservative approached was followed, and only the ESTs called out as significant by both statistical packages were considered for the subsequent analyses.
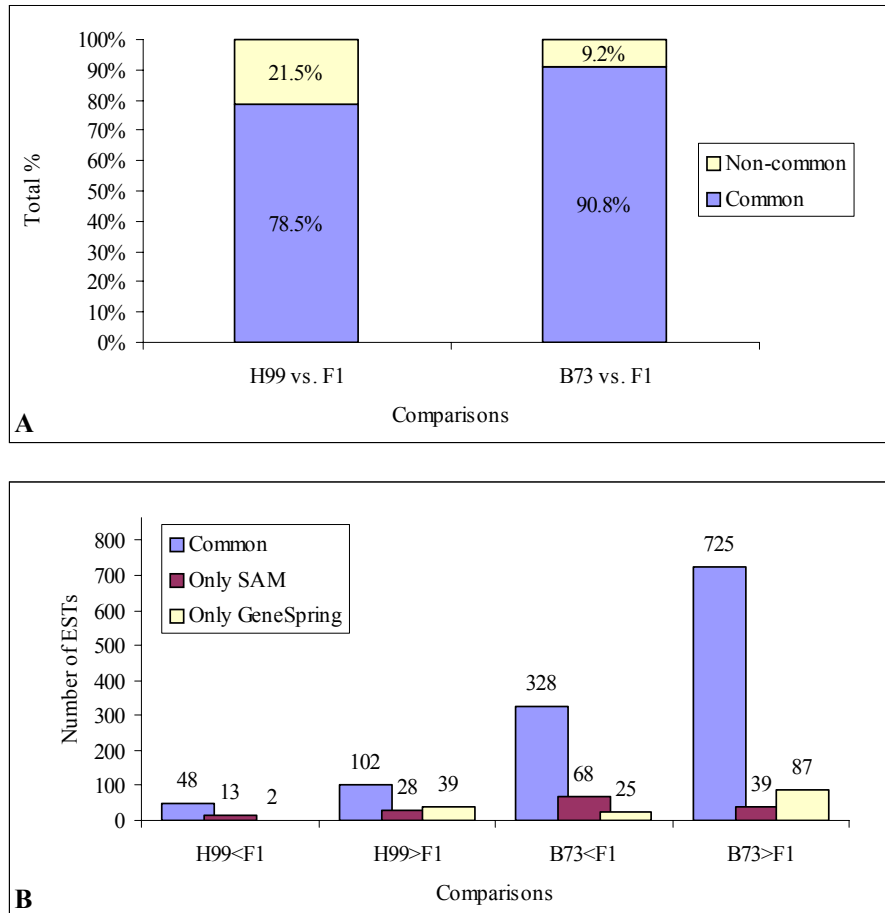
Figure 3: Comparison between GeneSpring and SAM significance analyses on microarray data. **A:** relative proportion of the ESTs called out by both significance analyses in H99 vs. F_1 and B73 vs. F_1 comparisons; **B:** histogram representing the number of significantly up and down-regulated ESTs for each comparison as they were called out by GeneSpring and SAM software (Common: ESTs called out by both significance analyses; Only SAM and Only GeneSpring: ESTs called out only by SAM or GeneSpring significance analysis respectively).

The proportion between up and down-regulated ESTs was similar in both H99 vs. F_1 and B73 vs. F_1 comparisons, showing a prevalence of ESTs up-regulated in the single parental lines (68.00% and 68.85% of the total regulated genes within H99 vs. F_1 and B73 vs. F_1 respectively). A number of ESTs shared similar regulation pattern in both inbred lines with the respect to the hybrid. Interestingly, in this case the relative proportion between up and down-regulated genes was inverted, with a prevalence (86.49%) of ESTs down-regulated in parental lines (Figure 4).
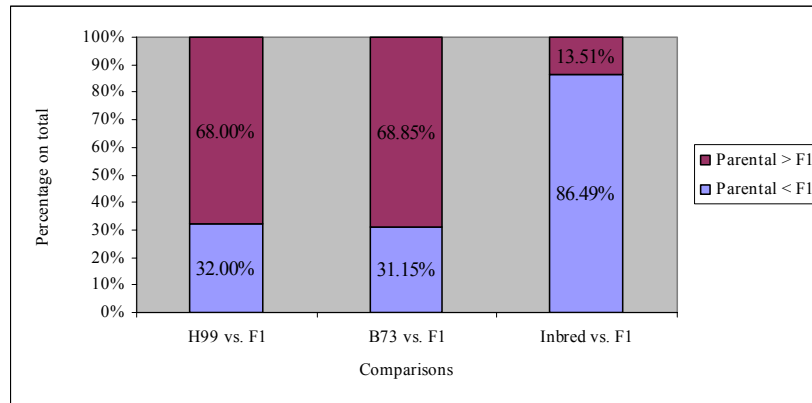
Figure 4: Relative distributions of up and down-regulated ESTs in the single inbred lines comparisons (H99 vs. F_1 and B73 vs. F_1) and the inbred-specific ones (inferred data). Significant ESTs commonly called out by GeneSpring and SAM software has been considered.

Ontological analysis

The ESTs detected as significant by both the software were assigned to Tentative Contigs at the TIGR database (ZM Gene Index, release 13.0). Their putative functions were then determined on the base of the Molecular Function categories reported in the TIGR Maize Gene Ontology database (http://www.tigr.org, Table 2).

| Maize Gene Ontology Categories	H99<F1		H99>F1		B73<F1		B73>F1		Inbred<F1		Inbred>F1	
	Hits	%	Hits	%	Hits	%	Hits	%	Hits	%	Hits	%
Catalytic	6	12.5%	15	14.7%	51	15.5%	124	17.1%	3	9%	1	20%
Transcription regulator	2	4.2%	2	2.0%	4	1.2%	48	6.6%	0	0%	0	0%
Translation regulator	4	8.3%	1	1.0%	12	3.7%	20	2.8%	3	9%	0	0%
Signal transducer	1	2.1%	1	1.0%	4	1.2%	11	1.5%	0	0%	0	0%
Chaperone	1	2.1%	1	1.0%	23	7.0%	30	4.1%	1	3%	0	0%
Antioxidant activity	0	0.0%	0	0.0%	0	0.0%	2	0.3%	0	0%	0	0%
Apoptosis regulator activity	0	0.0%	0	0.0%	0	0.0%	1	0.1%	0	0%	0	0%
Binding	4	8.3%	2	2.0%	1	0.3%	48	6.6%	0	0%	0	0%
Defense/immunity protein activity	0	0.0%	17	16.7%	2	0.6%	4	0.6%	1	3%	0	0%
Enzyme regulator activity	0	0.0%	1	1.0%	0	0.0%	3	0.4%	0	0%	0	0%
Motor activity	0	0.0%	0	0.0%	1	0.3%	0	0.0%	0	0%	0	0%
Nutrient reservoir activity	0	0.0%	0	0.0%	1	0.3%	2	0.3%	0	0%	0	0%
Obsolete	0	0.0%	0	0.0%	0	0.0%	2	0.3%	0	0%	0	0%
Other	0	0.0%	4	3.9%	18	5.5%	33	4.6%	2	6%	0	0%
Protein tagging	3	6.3%	0	0.0%	5	1.5%	10	1.4%	1	3%	0	0%
Structural molecule activity	0	0.0%	23	22.5%	50	15.2%	125	17.2%	3	9%	1	20%
Transporter	1	2.1%	2	2.0%	11	3.4%	17	2.3%	1	3%	0	0%
Unknown	26	54.2%	33	32.4%	145	44.2%	245	33.8%	17	53%	3	60%
Total	**48**	**100%**	**102**	**100%**	**328**	**100%**	**725**	**100%**	**32**	**100%**	**5**	**100%**

Table 2: Summary of the functional classification of the differentially expressed ESTs (TIGR ZM Gene Ontology).

When grouped among the functional categories, significant genes show comparable absolute distributions in H99 vs. F_1 and B73 vs. F_1, with the exception of *Defense protein activity* which are consistently more represented in H99 vs. F_1. *Catalytic* and *Structural*

molecule activity are the most represented single categories in both comparisons (Figure 5A). Up to 39.3% and 37.0% of significant genes could not be assigned to any known function in H99 vs. F_1 and B73 vs. F_1 respectively.

Relative distribution of up and down-regulated genes within each comparison is shown in Figure 5B. Among single functional categories, *Catalytic, Defense protein activity* and *Structural molecule activity* maintain the same trend in both comparisons, showing a prevalence of up-regulated genes in parental line; this is particularly evident for *Defense protein activity* and *Structural molecule activity* for which the totality of significant ESTs are up-regulated in H99 vs. F_1. A trend inversion in H99 vs. F_1 and B73 vs. F_1 could instead be observed within *Translation regulator activity*, which is also the only category showing a predominance of ESTs down-regulated in parental line (H99 vs. F_1). *Transcription regulator, Signal transducer* and *Chaperone* are equally distributed between up and down-regulated within H99 vs. F_1 comparison, while are preferentially up-regulated in parental line within B73 vs. F_1.
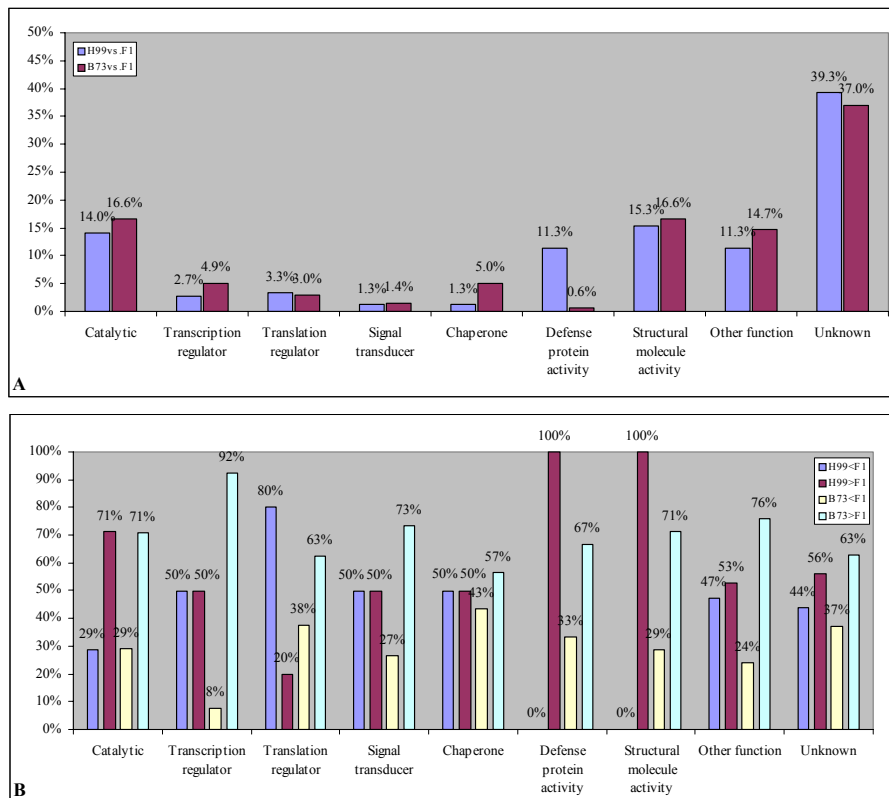
Figure 5: Histograms representing the proportion of the represented gene functions among the regulated genes in the two tested comparisons. **A**: absolute abundance of the significant genes within each comparison; **B**: relative distribution of the up and down-regulated genes within each comparison. For simplicity, only main functional categories are reported, while the other are grouped as "*Other function*"; genes for which the putative function is not known are categorized as "*Unknown*".

31

Real-time PCR

In order to get an estimate of the reliability of datasets obtained from the combined statistical approach on microarray data, the absolute expression level of a subset of ESTs were measured in all genotypes by real-time PCR. For this purpose, 3 ESTs assigned to each class of expression by statistical analysis of microarray data (H99<F_1, H99>F_1, H99=F_1 and B73<F_1, B73>F_1, B73=F_1) were sampled (Table 3). ESTs were chosen among those commonly called out as significant by both statistical analyses. Relative expression levels were then calculated as the ratio of absolute transcript abundance in inbred lines vs. hybrid, and compared to microarray data. 10 of the 18 tested ESTs confirmed the microarray results (Figure 6).

Expression category	Accession	Microarray (inbred/hybrid ratio)		RealTime (inbred/hybrid ratio)	
		Mean	SD	Mean	SD
H99<F1	AI881783	0.760	0.132	0.753	0.052
	AI666083	0.764	0.086	2.273	0.049
	AI881507	0.773	0.134	0.935	0.091
H99>F1	AI737795	1.402	0.188	2.150	0.099
	AI691932	1.412	0.041	1.375	0.007
	AI881226	1.422	0.103	3.670	1.895
H99=F1	AI665922	0.995	0.052	1.540	0.424
	AI770902	1.002	0.033	1.705	1.266
	AI714420	0.994	0.039	1.125	0.742
B73<F1	AI739775	0.343	0.053	0.680	0.071
	AI714512	0.380	0.029	0.900	0.269
	AI714507	0.394	0.080	0.790	0.127
B73>F1	AI734743	1.513	0.214	1.665	0.375
	AI734427	1.519	0.313	3.185	3.231
	AI881281	1.508	0.221	0.390	0.410
B73=F1	AI881808	1.003	0.051	1.245	0.502
	AI737778	1.005	0.036	0.675	0.262
	AI855088	1.006	0.058	1.530	0.240

Table 3: Summary of microarray vs. real-time PCR result comparison. Expression categories are defined according to statistical analysis of microarray data. Mean and standard deviation values are calculated on 5 and 8 replicates for microarray and real time PCR respectively.
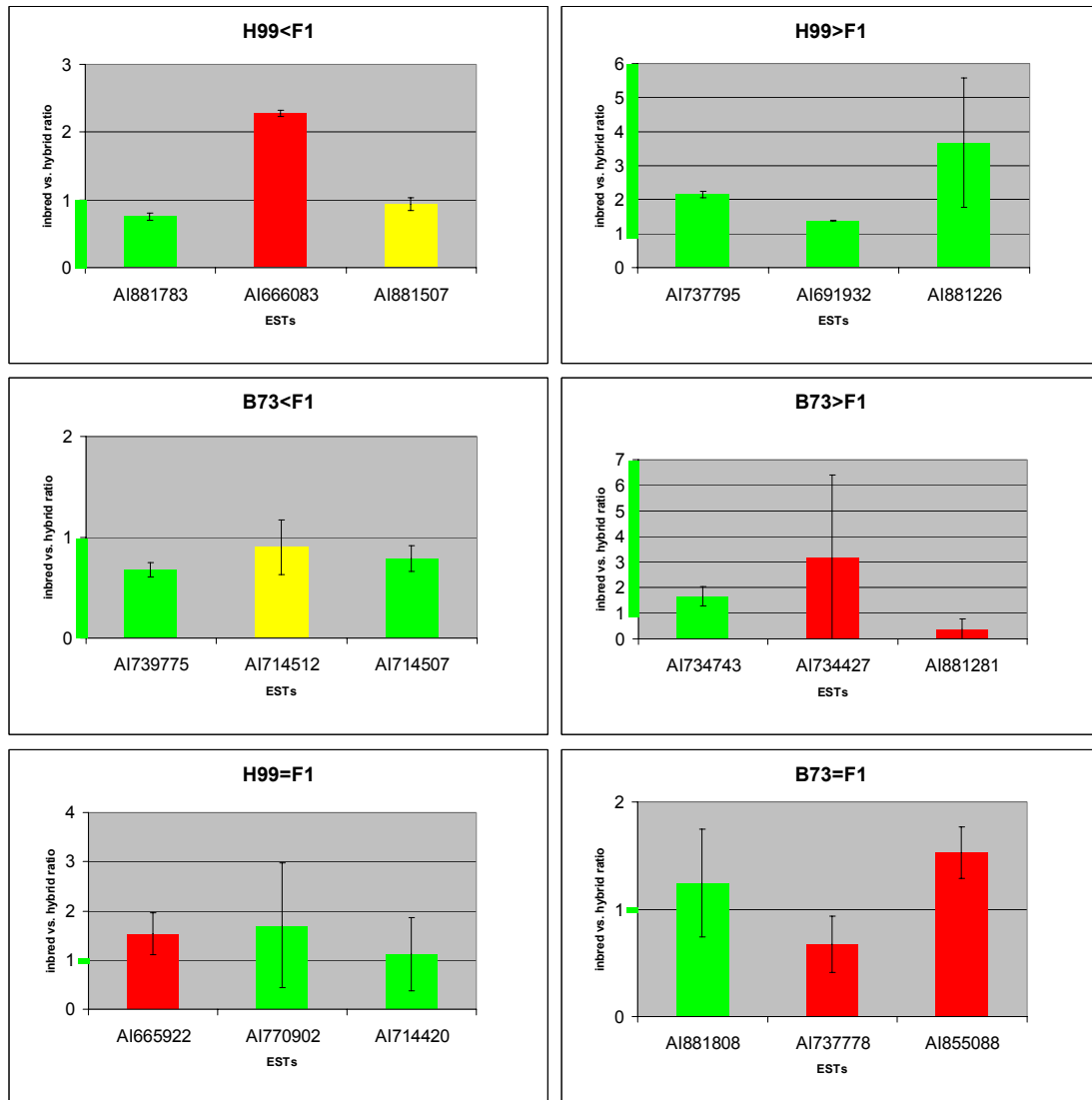
Figure 6: Histograms showing Real time PCR results. Each graph reports data for ESTs assigned to different category of expression by microarray statistical analysis; the green bar on the Y axis indicates the range of ratio values confirming microarray data (up-, down- on no expression regulation). Bars represent inbred line vs. hybrid expression ratios resulting from Real time PCR analysis (averaged values and standard deviation from replicates within the two best experiments, see Table 3); green and red colors indicate respectively confirmed and not confirmed results; yellow bars indicates non-conclusive data.

Discussion

In the literature, studies are reported showing differential gene expression between parental inbred lines and their corresponding hybrid (Romagnoli et al. 1990; Leonardi et al. 1991; Osborn et al. 2003; Song and Messing 2003). It was also proposed that differential gene expression could substantially contribute to hybrid vigor (Birchler et al. 2003). The question then becomes whether these changes are responsible for heterosis or a result of it; and if they are responsible, what property of heterozygousity would produce this response

at the target genes. Those studies relied on analyzing the expression of a few sampled genes, rather than on a comprehensive examination of wide expression patterns.

It has been recently pointed out that microarray technology might properly be applied to quantitative traits (Kerr and Churchill 2001; Wolfinger et al. 2001). We set out to assess the relative transcriptional levels of about 5,000 ESTs in immature ears of two parental inbred lines (B73 and H99) and their heterotic F_1 hybrid employing cDNA microarray technology. The goal was to survey transcription regulation on a large sample of genes in order to determine whether heterosis might be correlated with generalized changes in gene expression and/or with specific regulation of metabolic and regulatory patterns. To be taken in account is the fact that, even if our analysis can rely on a representative number of genes, it can not be regarded as comprehensive of the whole maize transcriptome. In fact, estimating from Lynx MPSS™ data (Brenner et al. 2000; Reinartz et al. 2002) that a total of about 14,000 genes are expressed in maize immature ear (B73 and Mo17 immature ear libraries; cut-off: 10ppm; M. Hanafey, personal communication), the ca. 1,700 different genes on the microarrays here used represent only about 12% of total transcripts.

In both H99 vs. F_1 and B73 vs. F_1 comparisons, we generally observed little differences in expression levels: substantially none of the expression ratio exceeded a ±2 fold-change, with most values assessed even below ±1.5 fold-change. However, multiple replicates carried out for each comparison allowed us to perform an efficient statistical data analysis, allowing us to identify sets of genes differentially expressed in terms of statistical significance. In fact, despite small differences in expression levels, by applying a robust statistical approach to microarray data, we were able to detect a large number of regulated genes between single parental lines and the hybrid. In order to increase the robustness of our survey, we also applied two different and independent statistical methods. Their outcomes were for large part comparable, confirming the reliability of statistic approach as it applies to microarray data. Significantly, 2/3 of ESTs within the tested subset were assigned by real time PCR to the class of expression determined by statistical analysis of microarray data, further confirming our results.

It is important to underline here that heterosis is a complex phenomenon. Therefore allelic interaction of many genes, rather than of single or a few genes, is likely to determine the heterotic phenotype. DNA microarray in transcriptome analysis derives most of its power in the identification of a small subset of genes for further characterization. This technology turns to be effective especially comparing treated vs. non-treated samples, or

wild-type vs. mutants, circumstances for which differences in expression are expected to be quite relevant. However, a challenging technical consideration in DNA microarray analysis is the cutoff value used to distinguish differential expression from natural variability in the data. A cutoff of twofold up- or down-regulation has been chosen to define differential expression in most published studies, but little has been done to evaluate the accuracy of the technique and assess the confidence levels of the two-fold level changes in expression ratios (Yang et al. 2002). In addition, the arbitrary choice of ratio thresholds has no statistical basis, and such approach does not provide the necessary flexibility for the analysis of complex traits. Conversely, the identification of statistical thresholds for the establishment of high-confidence sets of genes provides indubitably a better alternative in term of reliability/adaptability of the outcomes (Jin et al. 2001). This approach is probably the only applicable for investigating complex phenotypes on a large scale at the transcriptional level, for which changes far below the commonly accepted arbitrary cut-offs, but nonetheless biologically relevant, are likely to occur (Jin et al. 2001). This fully applies to such a puzzling and hidden phenomenon as heterosis.

Microarray data show that H99 is more similar in gene expression to the hybrid than B73 (only 3.7% of the analyzed ESTs are differentially regulated in H99 vs. F_1, against 25.7% in B73 vs. F_1). This is surprising, considering the phenotypic characteristics of the two parental lines as compared to their hybrid. In fact, a quantitative survey we recently conducted showed that B73 is the best parent in term of yield and yield components traits, i.e. B73 is the parental line performing closer to the B73 x H99 heterotic hybrid. This indicates that, at least in the developing ear, the simple overall amount of expression difference can not be correlated to the heterotic phenotype.

Most of the significant genes are regulated only in one or the other parental line, or in other terms, each of the parental lines differs from the hybrid in a unique way. This suggests that the overall phenotypic effect observed in the hybrid depends on how genes are differently affected in their expression level in heterozygous background, rather than on the absolute number of genes differentially regulated between inbred lines and hybrid. The presence of both up-regulated and down-regulated genes further indicates that heterosis might not be related to a simple increase or decrease of transcript levels in particular genes, but rather to the occurrence of peculiar regulative combinations due to the heterozygous state of the hybrid. However, both in H99 vs. F_1 and B73 vs. F_1 about 68% of the significant ESTs result up-regulated in parental lines, indicating for most significant genes a major trend to be down-regulated in the hybrid. A few genes commonly regulated

in both the parental lines with respect to the hybrid were also observed. Noticeably, for these genes a prevalence of down-regulation in parental lines was observed (86.49%). This might indicate that a number of genes may also exist which are subjected to a specific regulation mechanism, mainly consisting in a release-of-repression within the hybrid, which also may be of significance for heterosis.

Significant genes observed in this study are related to a range of structural, enzymatic and regulatory functions, with various possible sub-cellular localizations. This supports the hypothesis that hybrid vigor is due to gene expression regulation encompassing a broad range of biochemical pathways rather than definite components within localized processes. It would in fact be reasonable to assume that heterosis, involving tiny but wide-ranging differences, could arise from extensive genomic-environmental regulative pathways acting on the whole phenotype through some pleiotropic regulation system. Furthermore, inbred lines show a higher level of expression than the hybrid for most of the regulated genes within functional categories. This might suggest that heterozygotes are able to maintain optimal conditions with fewer metabolic efforts, with a positive effect on the overall performance. This "hybrid advantage" might come either from the possibility for a heterozygote to "choose" the best parental allele (dominance) or the best allele combination (over-dominance), or both. Further, each combination can be established in an adaptable way in different tissues, environments or developmental stages, either within the same gene or among different genes (epistasis), and either within the same metabolic pathways or among different ones. To be noted is that such wide-spread regulatory mechanism does not require dramatic changes in expression levels in order to greatly affect the overall plant phenotype.

Unfortunately, all molecular and genetic evidences could both be seen as underlying heterosis or as direct effects of its establishment. Furthermore, the task of assigning a functional meaning to results is also complicated by the fact that more than 1/3 of significant genes could not be assigned to any known function, which also underlines the actual limits of maize sequences annotation. Finally, due to complexity of interactions, it is likely that different causes/effects are produced depending on the particular feature or component under study. All this might have probably led in the past to some of the contradictory interpretations of hybrid vigor, and might explain why precise mechanisms most intimately involved in controlling heterosis still remain uncovered.

However, our study accomplished to combine microarray technology and statistics for the analysis of a complex trait. Despite the small differences observed in transcription

levels, by applying a robust statistical analysis, we succeeded in individuating significantly regulated genes that might be correlated to heterosis and thus provide the bases for further analyses. Our data indicate a prevalent multigenic nature of heterosis at the transcriptional level, affecting various cellular and molecular functions. However, also indications that genes specifically regulated in the hybrid and thus expressly contributing to hybrid vigor might exist were produced. We were also able to exclude the existence of a simple correlation between the overall variability in gene expression and the heterotic phenotype, which in turn resulted mainly influenced by the quality of transcripts, as well as by the establishment of peculiar regulative interactions within hybrids, which not necessarily imply dramatic changes in gene expression. Finally, lists of genes potentially involved in heterosis have been produced, for which *cis*- and *trans*-regulatory elements might be investigated for presence of motives commonly influenced when in heterozygous status.

Materials and methods

Plant material

Plant material was collected from maize inbred lines B73 and H99, as well as from their F_1 hybrid (B73 x H99). Immature ears were harvested from plants cultivated in open field, selecting those whose silks reached no more than two third of the ear length. Material was immediately frozen after removing silks and ear apexes. To minimize environmental and individual variances, the plant material was collected at the same time of the day. All plant material was stored at -80°C until RNA extraction.

cDNA microarray

cDNA microarrays (print n° 606.01.04) produced at the University of Arizona, Tucson, were used to investigate expression levels of B73 and H99 maize inbred lines versus their F_1 hybrid. A total of 15,606 DNA are spotted in triplicate on each glass-slides, representing 4906 Zea mays expressed sequence tags (ESTs) from immature ear tissue cDNA library (Schmidt lab, UCSD) and 111 different controls (details on array format available at http://www.maizegdb.org/documentation/mgdp/microarray/).

Total RNA isolation and poly(A^+) purification

To minimize individual differences in transcript levels, material from at least 10 different plants for each genotype was bulked and melted together prior to extraction. Tissues were grinded in liquid nitrogen using mortars and pestles. Total RNA was isolation

using TRIZOL protocol (Invitrogen, Carlsbad, CA), as indicated by the manufacturer (except for 5 minutes extra time centrifugation in TRIZOL reagent), including a second step in chloroform for lower protein contamination. Total RNA was resuspended in DEPC-treated mQ water and stored at -80°C. For long storage and transport, RNA was precipitated in 0.3M NaCl and 2.5 volumes chilled 100% ethanol (for later use, RNA was resuspended in DEPC-treated mQ water, after 20 minutes centrifugation at 4°C). Poly(A$^+$) RNA was purified from 1mg of total RNA using mRNA Purification Kit (Amersham Bioscience, Little Chalfont, UK). Both total and Poly(A$^+$) RNA have was tested for quality by electrophoresis on 1.5% agarose gel and quantified by absorbance at 260 nm.

Array hybridization and fluorescence detection

1μg of purified poly(A$^+$) RNA from each genotype was retrotranscribed using 400U of SuperScript II RNase H$^-$ Reverse Transcriptase (Invitrogen, Carlsbad, CA) and 2μg Oligo(dT)^{23} Anchored (Sigma-Aldrich, St. Louis, MO) as primer, in 30μl final volume (2h, 42°C). cDNA probes were labeled by direct incorporation of Cy3-Cy5 modified dCTP, final concentration 0,3 mM (Amersham Bioscience, Little Chalfont, UK) during retrotranscription (dATP, dGTP e dTTP 0,5 mM each, dCTP 0,2 mM). Reaction was blocked adding 1.5μl EDTA (0,5M - pH8) and 3.75μl NaOH (1M) (10 min., 65°C) and then neutralized with 0.75μl HCl (5M) and 9μl Tris HCl (1M - pH6.9). Probe was purified with Nucleo Spin Extract kit (MACHEREY-NAGEL GmbH & Co. KG, Düren, Germany), protocol # 4.2 with double wash in NT3 buffer. After adding 12μg of Polydeoxyadenylic Acid (Amersham Bioscience, Little Chalfont, UK) the probe was lyophilized in SpeedVac™ SVC-100 H (Savant Instruments/E-C Apparatus, Holbrook, NY) and then resuspended in 29μl Array Hyb Low Temp Hybridization Buffer (Sigma-Aldrich, St. Louis, MO) and 2μl salmon sperm DNA (20μg/μl). After denaturation (2min., 98°C) probe was hybridized o.n. at 50°C on microarray slides in hybridization chamber (CMT-Hybridization Chamber, Corning Inc., Corning, NY). After washings, microarray images were acquired by ScanArray® v3.1 software on SA4000 Scanner (Packard BioScience, Wellesley MA). Spot fluorescence intensity was quantified by QuantArray® v3.0 software (Packard BioScience, Wellesley MA). Single bad-quality spots were individuated by hand and flagged in order to be taken in account in further steps of analysis.

Microarray data normalization and statistical analysis

Raw data were normalized using GeneSpring software (Silicon Genetics, Redwood City, CA - demo license). First of all, a reference "genome" was created, including the list of all the ESTs spotted on the immature ear microarray. Before data loading, raw intensity records were manually edited and data relative to bad-quality spots in each hybridization were deleted. The LOWESS intensity dependent normalization ("non-linear" normalization) was firstly applied to data, in order to correct for artifacts caused by non-linear rates of dye incorporation as well as inconsistencies in the relative fluorescence intensity between dyes. Per Chip normalization ("Normalize to a median or percentile" function) was subsequently applied, in order to correct for chip-wide variations in intensity that may have been due to inconsistent washing, inconsistent sample preparation, or other microarray production or micro-fluidics imperfections. A 50th percentile value was used, with "extra background correction if necessary" option on. For statistical analysis, normalized data, exported as single EST averaged ratio in spreadsheets from GeneSpring, were transformed into $\log_2$ and was submitted to one-class response format in SAM software (Tusher et al. 2001), considering here each microarray experiment as a replica. ESTs with more than 3 missing experiments were deleted before running the program. One class response analysis was used, submitting each ratio value into the SAM software as the $\log_2$(signal/control) ratio. A validation of SAM called-out statistically significant genes was conducted by GeneSpring filtering on confidence function. T-test filtering, corrected for multiple samples (Bonferroni and Hochberg false discovery rate), was used. P-values were set so that the output presenting the same number of significant ESTs detected by SAM according to what described above.

Real-time PCR

The primer sets for real-time PCR (optimum length 20bp; Tm 60°C; GC% ≥55%; Table 4) were designed to the sequence of each of the tested ESTs using Primer3 software (Rozen and Skaletsky 2000). Amplification products of 150-200bp, as close as possible to the 3'-end of sequences, were chosen. ESTs sequences were also checked for the presence of secondary structures possibly impeding the real-time reactions, using Mfold software (Zuker 2003) with a cut-off of ΔGo>-6kcal/mol. For the real-time PCR assay, total RNA was treated with Deoxyribonuclease I (DNase I Amplification Grade, Sigma-Aldrich, Saint Louis, MO - Cat No. AMP-D1) as reported by manufacturer, dried by vacuum speed centrifugation (SpeedVac™ SVC-100 H, Savant Instruments/E-C Apparatus, Holbrook,

NY) and resuspended in 20µl DEPC water. As internal control, an aliquot of zebrafish (*Danio rerio*) mRNA for *otx homeoprotein 3* (*OTX3*) was added to each RNA sample prior to retrotranscription, corresponding to 1/20,000 of total RNA amount; primers were designed to the first 600bp of the *OTX3* complete mRNA sequence (gi:633134), which showed no homologies with any maize sequences in TIGR and GenBank databases (blastn alignment). cDNA synthesis was then carried out by iScript cDNA Synthesis kit (Bio-Rad Laboratories, Hercules, CA – Cat No. 170-8890), with a 60 min. reaction time; at the end, each sample was diluted to 100ng/µl. Each Real-time PCR reaction were carried out on 380ng of cDNA, using the iQ SYBR Green 2X Supermix Kit (Bio-Rad Laboratories, Hercules, CA – Cat No. 170-8882) in a 25µl total volume. Real time PCR was performed on ICycle thermo cycler (mod. SBI002.0, Bio-Rad Laboratories, Hercules, CA). Amplification cycles: 3 min., 95°C; 46 cycles 30 sec. 95°C, 40 sec. 60°C, 40 sec. 72°C; 10 min. 72°C; melting curve: from 55°C to 95°C, +0.5°C increment at each cycle. 3 separately retrotranscribed cDNA were tested (replicates), performing 4 measurements each in the same reaction plate. Titration curve was built on *OTX3* signal in 4 serial dilution of the template (1:1, 1:10, 1:100 and 1:1,000). All the reaction plates also included positive (*actin*) and negative (no cDNA) control.

Class	Accession	Forward Primer (5′-3′)	Reverse Primer (5′-3′)
	AI881783	GATATCGGTGCTCCCTTGAC	CCTCTGCTTGGAATTGCTG
H99<F1	AI666083	GCCAATACAAGCGGGTAGAC	AGCTAGACGAGTGCGAGGAG
	AI881507	TGAGCTTCCGAGTAGTTCAGG	ATCCCCGTCCTTCTACTGGT
	AI737795	GGGACACTCATCACCACAGA	CATCGTGCTCTGGAAGTGG
H99>F1	AI691932	TCGACCCTCACTTCTCTTGG	TACCATCACCATCGGCATC
	AI881226	CATCGTGCTCTGGAAGTGG	GGGACACTCATCACCACAGA
	AI665922	CCTGCAGGCAACATAGCAT	CCTGTGGTGTACCTGTTTCG
H99=F1	AI770902	AGGAACGTGCAGGCGAAT	AGCTAGCGCTGCTCTCCA
	AI714420	CCACCATGTATGAGGGGAAC	GGAAGGTGCTCAAGTGGAAG
	AI739775	ACGAGATCGCTTCACACCTC	GTACGAGAGGACTGGGTTGG
B73<F1	AI714512	ACGAGATCGCTTCACACCTC	GTACGAGAGGACTGGGTTGG
	AI714507	ACGAGATCGCTTCACACCTC	GTACGAGAGGACTGGGTTGG
	AI734743	GCTCTTGCCCTTCTTCCTCT	AGGAGGGACGTACCCTTGAC
B73>F1	AI734427	AGCCTCCACAGAGGTGATGT	GGAGCCATTCAAGGTGGTAG
	AI881281	TCGAGGATGGAGAGTGGTTC	CAAGGACGAGAGGCTGTAGG
	AI881808	GCTCGGAGACCTACAGCTTG	TTCCATGTTCTGGCCACTC
B73=F1	AI737778	AGGAGCCTTGAAGTGCTCTG	GCTGAAGCTTGGCATGAAG
	AI855088	GAGACCAGACACAGCAGCAC	AGAAGCCCACCATCACCTC

Table 4: List of real time PCR primer sets

References

Barbosa, A.M.M., Geraldi, I.O., Garcia, A.A.F., Souza, C.L. and Souza, A.P. (2003). "Relationship of intra- and interpopulation tropical maize single cross hybrid performance and genetic distances computed from AFLP and SSR markers." Euphytica **130**: 87-89.

Birchler, J.A., Auger, D.L. and Riddle, N.C. (2003). "In search of the molecular basis of Heterosis." The Plant Cell **15**: 2236-2239.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G., Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J. and Corcoran, K. (2000). "Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays." Nat Biotechnol **18**(6): 630-634.

Comings, D.E. and MacMurray, J.P. (2000). "Molecular heterosis: a review." Mol Genet Metab **71**: 19-31.

Coors, J.G. and Pandey, S. (1999). Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA.

Davenport, C.B. (1908). "Degeneration, albinism and inbreeding." Science **28**: 454-455.

DeVienne, D., Bost, B., Fievet, J., Zivy, M. and Dillmann, C. (2001). "Genetic variability of proteome expression and metabolic control." Plant Physiol Biochem **39**: 271-283.

Dobzhansky, T. (1950). "Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of Drosophila pseudobscura." Genetics **35**: 288-302.

Duvick, D.N. (2001). "Biotechnology in the 1930s: the development of hybrid maize." Nat Rev Genet **2**(1): 69-74.

East, E.M. (1908). "Inbreeding in corn." Rep Connecticut Agric Exp Stn for 1907**:** 419-428.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A. and Smith, O.S. (2004). "Allelic Variation of Gene Expression in Maize Hybrids." Plant Cell **16**(7): 1707-1716.

Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G. and Gibson, G. (2001). "The contribution of sex, genotype and age to transcriptional variance in Drosophila melanogaster." Nat Genet **29**: 389-395.

Kerr, M.K. and Churchill, G.A. (2001). "Statistical design and the analysis of gene expression microarray data." Genet Res **77**: 123-128.

Leonardi, A., Damerval, C., Herbert, Y., Gallais, A. and DeVienne, D. (1991). "Association of protein amount polymorphisms (PAP) among maize lines with performances of their hybrids." Theor Appl Genet **82**: 552-560.

Liu, X.C., Ishiki, K. and Wang, W.X. (2002). "Identification of AFLP markers favorable to heterosis in hybrid rice." <u>Breed Sci</u> **52**: 201-206.

Melchinger, A.E. (1999). Genetic diversity and heterosis. <u>In: Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA</u>**:** pp 99-118.

Moll, R.H., Lonnquist, J.H., Fortuna, J.V. and Johnson, C.E. (1965). "The relationship of heterosis and genetic divergence in maize." <u>Genetics</u> **52**: 139-144.

Osborn, T.C., Pires, J.C., Birchler, J.A., Auger, D.L., Chen, Z.J., Lee, H.S., Comai, L., Madlung, A., Doerge, R.W., Colot, V. and Martienssen, R.A. (2003). "Understanding mechanisms of novel gene expression in polyploids." <u>Trends Genet</u> **19**: 141-147.

Reinartz, J., Bruyns, E., Lin, J.Z., Burcham, T., Brenner, S., Bowen, B., Kramer, M. and Woychik, R. (2002). "Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms." <u>Brief Funct Genomic Proteomic</u> **1**(1): 95-104.

Romagnoli, S., Maddaloni, M., Livini, C. and Motto, M. (1990). "Relationship between gene expression and hybrid vigor in primary root tips of young maize (Zea mays L.) plantlets." <u>Theor Appl Genet</u> **80**: 769-775.

Rozen, S. and Skaletsky, H.J. (2000). Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ**:** 365-386.

Shull, G.H. (1908). "The composition of a field of maize." <u>Rep Am Breeders Assoc</u> **4**: 296-301.

Song, R. and Messing, J. (2003). "Gene expression of a gene family in maize based on non-collinear haplotypes." <u>Proc Natl Acad Sci USA</u> **100**: 9055-9060.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." <u>Proc Natl Acad Sci USA</u> **98**: 5116–5121.

Wolfinger, R.D., Gibson, G., Wolfinger, E.D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C. and Paules, R.S. (2001). "Assessing gene significance from cDNA microarray expression data via mixed models." <u>J Comput Biol</u> **8**(6): 625-637.

Yang, I.V., Chen, E., Hasseman, J.P., Liang, W., Frank, B.C., Wang, S., Sharov, V., Saeed, A.I., White, J., Li, J., Lee, N.H., Yeatman, T.J. and Quackenbush, J. (2002). "Within the fold: assessing differential expression measures and reproducibility in microarray assays." <u>Genome Biol</u> **3**(11): research0062.1–0062.12.

Yang, Y.H. and Speed, T. (2002). "Design issues for cDNA microarray experiments." <u>Nat Rev Genet</u> **3**: 579-588.

Zuker, M. (2003). "Mfold web server for nucleic acid folding and hybridization prediction." <u>Nucl. Acids. Res.</u> **31**(13): 3406-3415.

Expression analysis of allele-specific genic insertions in maize (*Zea mays* L.)

Giorgio Pea[1,2], Stephan Brunner[1], M. Enrico Pè[2], Michele Morgante[3], and Antoni Rafalski[1]

[1]DuPont Crop Genetics Research, DuPont Experimental Station Building E353, Wilmington, DE19880-353, USA; [2]Università degli Studi di Milano, Dipartimento di Scienze Biomolecolari e Biotecnologie – via Celoria 26, 20133 Milano, Italy; [3]Universita' degli Studi di Udine, Dipartimento di Scienze Agrarie ed Ambientali - via delle Scienze 208, 33100 Udine, Italy

Abstract

Lack of intraspecific micro-collinearity has been reported in maize at two different loci (Fu and Dooner 2002; Song and Messing 2003). These results were recently confirmed from the comparison of four additional large allelic genomic sequences between the inbred lines B73 and Mo17 (Brunner et al. submitted). Among the non-shared sequences, mainly represented by repetitive element insertions, genic sequences were also found; these frequently corresponded to gene fragments organized in exon-intron structures, with the same orientation and arranged within clusters and were referred as *non-shared genes*.

We set out to investigate the genomic organization and the expression of some of the non-shared gene clusters within three of the genomic contigs compared by Brunner and coworkers (contigs 9002, 9008 and 9009). PCR assays on genomic DNA showed that sequences homologous to the single non-shared genes were present in both inbred lines. Analysis of oat-maize addition lines using the same PCR primer sets allowed us to estimate that these sequences were present in multiple copies in the genome and to assign them to individual maize chromosomes. The same analysis conducted with primer sets specific for the combinations of non-shared genes present as clusters within the compared contigs showed that such clusters were also present in other genomic locations in both lines, although usually in a lower number of copies. We then examined by RT-PCR the expression pattern of some of the non-shared genes as well as of their clusters, in different tissues collected from the B73 and Mo17 inbred lines as well as their reciprocal hybrids. In general, the mRNA corresponding to single genes was found by RT-PCR in all the tested samples. Interestingly, expression was also detected when testing primers designed to neighboring genes within clusters, indicating that they were actually transcribed as single

mRNA, and could possibly represent novel genes originating from modular shuffling of different gene fragments. Further, expression of some of these gene clusters was specific to one or the other of the two inbred lines, indicating differences in transcription regulations. This observation suggests that a gene expression pattern characteristic to a hybrid could originate from the combination of differentially regulated non-shared alleles present in the parental lines. Therefore, the parental gene sets, contributing differently to the transcriptome in crosses between different "non-homologous" partners, could contribute to the hybrid's attributes such as heterosis, which could neither be established nor fixed in a single homozygote line.

We discuss the possibility that the observed lack of collinearity in maize might contribute to heterosis, not only through the differences in transcription regulation between alleles (Guo et al. 2004), but also by creating a unique gene complement in the maize hybrid.

Introduction

Heterosis (or hybrid vigor) is defined as the superiority of a F_1 hybrid over its homozygote parental lines. It was first recognized and studied in maize (East 1908; Shull 1908) and it is extensively exploited in crop production. The understanding of heterosis could contribute greatly to the improvement of agronomically important traits. The genetic basis of heterosis has been extensively discussed in the past years, but so far little consensus has been reached (Lamkey and Edwards 1999). The classic quantitative genetic explanations for hybrid vigor include both the effect of dominance - i.e. complementation of deleterious alleles in the hybrid - and over-dominance - i.e. advantageous allelic interactions in the heterozygous background (Davenport 1908; East 1908; Shull 1908; Crow 1948). More recent applications of modern quantitative genetics have not been able to improve the understanding of the genetic nature of heterosis in maize as well as in other species (Stuber et al. 1992; Schnell et al. 1996; Yu et al. 1997; Meyer et al. 2004). Several molecular hypotheses about heterosis have also been integrated into the genetic models, but the phenomenon of heterosis still remain poorly understood (Birchler et al. 2003).

The analysis of genetic diversity among individuals of the same species is usually based on the assumption that an allelic counterpart is found for each gene. However, this assumption was challenged when two maize inbred lines, McC and B73, have been recently compared by over more than 100 kb at the *bronze1* genomic region (Fu and

Dooner 2002). Contrarily to expectations, the DNA sequences of the two inbred lines differ extensively in the repetitive DNA segments, and genes present in one allele were also found missing in the relative counterpart. Similar results were reported at the maize *c1Z* cluster locus (Song and Messing 2003). To estimating the frequency of occurrence and the extent of this phenomenon in the maize genome, comparative sequence analyses have been recently extended to 4 additional collinear loci (ca. 250 kb each) in the B73 and Mo17 inbred lines (Brunner et al. submitted), as well as to the *bz1* Mo17 allele. Inter-genic non-homologies as well as differences in the gene content were found within all the analyzed regions. The confirmation of the previous observations suggests that lack of intraspecific micro-collinearity, involving both genic and non-genic elements, is indeed a common attribute of the maize genome. This phenomenon might account for the widespread allelic differences in gene expression observed in maize hybrids (Guo et al. 2004). These differences could be the results of different "sequence environments" surrounding the two alleles of a gene common to both alleles. One should also consider the possible contributions to the hybrid of additional non-shared gene sets present in each the parental inbreds (Fu and Dooner 2002). Both phenomena have been suggested as possible molecular explanations for hybrid vigor in maize (Birchler et al. 2003). Further, Brunner et al. found that the individual non-shared gene segments were commonly found arranged in clusters with other genic segments present in the same orientation. It has been proposed that introns and exons structures might have been involved in the modular formation of the first eukaryote genes (Doolittle 1978). Recombination within introns and exon shuffling might be responsible for the creation of novel complex gene functions from simple protein motifs by stochastic combination of small blocks of coding sequence whose product are then subjected to selection (Blake 1983) The peculiar clustering observed for the non-shared genes suggests that if they were transcribed as single mRNA, they might represent examples of intermediate steps in the creation of new genes functions by modular shuffling and combination of pre-existing functional motifs. A crucial question that remains to be answered is if these genic clusters are indeed transcriptionally active.

To contribute to the resolution of this issue, in this study we focused on the genomic organization and the transcriptional profile of some of these non-shared gene clusters; and on the presence of transcription across genes within the same clusters.

Results

Gene annotation and PCR primer design

Non-shared genes, specific to maize inbred lines B73 and Mo17 are frequently incomplete, but clustered and oriented in the same direction (Brunner et al. submitted). Here, we defined the exact Exon/Intron structure within and between the non-shared gene fragments at the loci 9002, 9008 and 9009 based on FGENSH splicing site predictions or blastx alignments to known proteins from the GenBank database and homologies to maize ESTs (Table 1). In some cases only fragments of single exons were detected (e.g. in locus 9002 genes G, H, I, J, N, K and L), whereas in many other situations (e.g. in locus 9002 genes O, P and Q) two or more Exons of a gene were present (Table 2). Specific PCR primer sets were then designed within single genes. Combinations of these primers were used to amplify across neighboring genes of non-shared genes clusters. Below, each locus is described in detail and summarized in Table 2.

Four non-shared gene clusters, present only in the B73 inbred line were considered on contig 9002: GHIJ9002, KLM9002, NOPQ9002 and RST9002. The cluster GHIJ9002, KLM9002 and RST9002 are oriented in *5'-3'* direction, relative to DNA sequence reported by Brunner et al (GenBank accession AY664413). The cluster GHIJ9002 covers a 6728bp region: gene G9002 and H9002 are separated by 569bp, while the gene pairs H9002-I9002 and I9002-J9002 are more distant from each other (2374bp and 3044bp respectively). In the cluster KLM9002 (3085bp), the genes are separated from each other by 701bp (K9002 to L9002) and 1379bp (L9002 to M9002). The four genes in the following cluster NOPQ9002 are within a 3442bp region. They are separated by 410bp (N9002 to O9002), 500bp (O9002 to P9002) and 305bp (P9002 to Q9002), respectively, and all are in *3'-5'* orientation relative to GenBank entry AY664413. The next cluster (RST9002) covers 6971bp and the genes are separated by 954bp (genes R9002 and S9002) and 1812bp (gene S9002 and T9002).
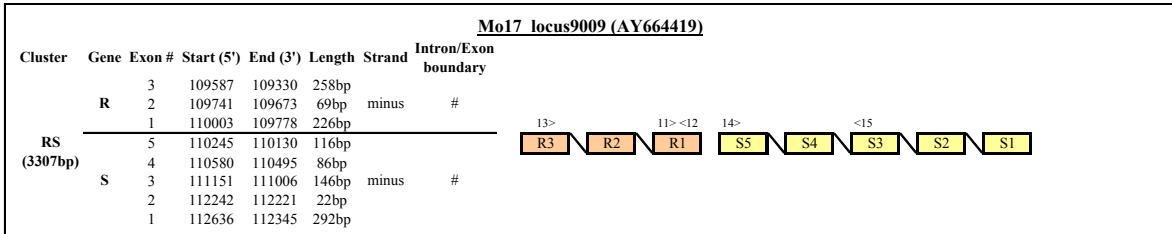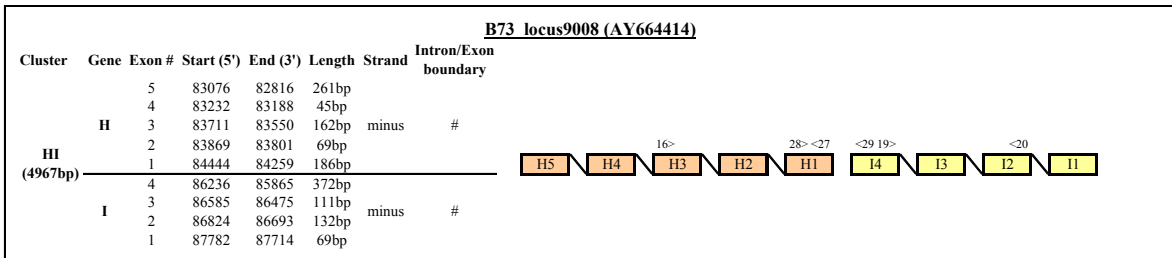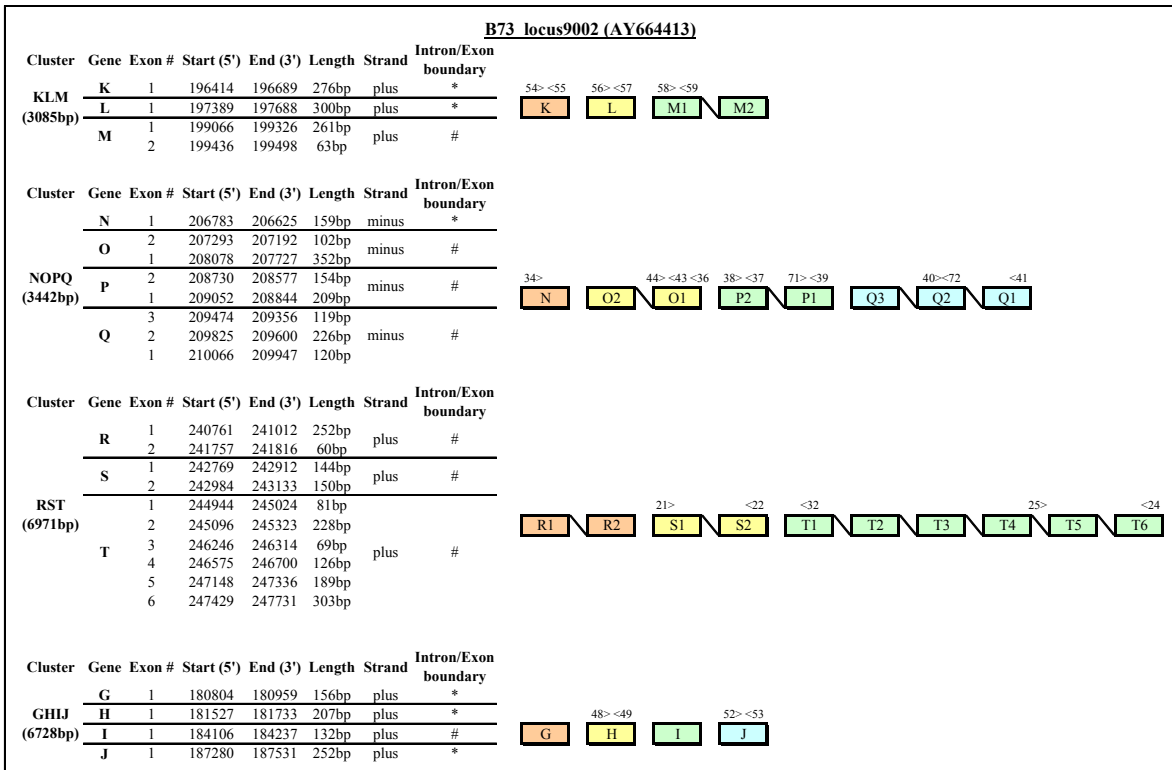
At locus 9008, the two genes of the non-shared cluster HI9008 (4967bp) are orientated in *3'-5'* direction relative to GenBank accession AY664414 and separated by 1422bp from each other.

Within contig 9009, only the Mo17 specific cluster RS9009 (3307bp) was analyzed, which consists of two genes in *3'-5'* direction relative to GenBank accession AY664419 and separated by 128bp.

gene nomenclature	locus	FGENESH prediction	genebank accession #	B73	Mo17	exon #	intron #	AA #	chain	Blast hit	E-value	rice: chromosome (PAC clone)	begin on rice	end on rice
geneG9002	9002	B1011A07.25 [Oryza sativa]	NP_908602.1	+	-	2	1	207	+	incomplete	5E-13	1S (AP003722)	-	-
geneH9002	9002	expressed protein (with alternative splicing) [Oryza sativa]	AAR87203.1	+	-	6	5	494	+	incomplete	8E-16	3L (AC090683)	-	-
geneI9002	9002	protein kinase family protein [Arabidopsis thaliana]	NP_187044.1	+	-	1	-	701	+	incomplete	4E-13	-	-	-
geneJ9002	9002	putative PRLI-interacting factor N [Oryza sativa]	NP_913434.1	+	-	1	-	529	+	incomplete	2E-10	1L (AP002902)	-	-
geneK9002	9002	OSJNBb0022F23.8 [Oryza sativa]	CAE02871.2	+	-	4	3	286	+	incomplete	5E-10	4L (AL606447)	-	-
geneL9002	9002	putative phosphatidylinositol/phosphatidylcholine transfer protein [Oryza sativa]	BAD07999.1	+	-	12	11	624	+	incomplete	0.00000003	2S (AP005851)	-	-
geneM9002	9002	putative AMP deaminase [Oryza sativa]	NP_910462.1	+	-	2	1	815	+	incomplete	2E-17	7L (AP004333)	-	-
geneN9002	9002	40S ribosomal protein S8 [Zea mays]	Q08069	+	-	1	-	53	-	incomplete	6E-21	2L (P0483C08)	-	-
geneO9002	9002	unknown protein [Oryza sativa]	BAC84209.1	+	-	2	1	159	-	incomplete	2E-41	7L (AP005259)	-	-
geneP9002	9002	unknown protein [Oryza sativa]	NP_915330	+	-	2	1	121	-	incomplete	3E-38	11S/12S	-	-
geneQ9002	9002	putative cytosolic monodehydroascorbate reductase [Oryza sativa]	AAL87167	+	-	3	2	153	-	incomplete	3E-28	2L (AF480496)	-	-
geneR9002	9002	putative hairpin inducing protein [Oryza sativa]	AAR88579.1	+	-	2	1	171	+	incomplete	2E-38	3L (AC092557)	-	-
geneS9002	9002	origin recognition complex subunit 1 [Zea mays]	AAL10452.1	+	-	4	3	202	+	incomplete	4E-28	6S (AQ869921)	-	-
geneT9002	9002	lysine-ketoglutarate reductase/saccharopine dehydrogenase bifunctional enzyme [Zea mays]	AAC18622.2	+	-	6	5	334	+	incomplete	2E-61	2L (AP004849)	-	-
gene nomen-clature	locus	FGENESH prediction	genebank accession #	B73	Mo17	exon #	intron #	AA #	chain	length of Blast hit	E-value	rice: chromo-some (PAC clone)	begin on rice	end on rice
geneH9008	9008	MADS box proten ZMM17 [Zea mays]	Q8VWM8	+	-	2	1	259	-	incomplete	5E-19	on several chromosomes	159,332	158,339
geneI9008	9008	putative phosphoinosititde phosphatase [Oryza sativa]	AAK92639.1	+	-	6	5	259	-	incomplete	4E-77	3S (AC079633)	111,990	112,265
gene nomen-clature	locus	FGENESH prediction	genebank accession #	B73	Mo17	exon #	intron #	AA #	chain	length of Blast hit	E-value	rice: chromo-some (PAC clone)	begin on rice	end on rice
geneR9009	9009	putative MAP3K epsilon protein kinase [Oryza sativa]	AAL87195.1	-	+	20	19	1264	-	incomplete	2E-35	4S (AL606608)	-	-
geneS9009	9009	putative splicing factor 3 [Oryza sativa]	AAO38832.1	-	+	5	4	587	-	full length	3E-11	3L (AC091532)	-	-

Table 1: Non-shared genes annotation results. Exons and Exon/Intron structures were identified either by FGENSH splicing site predictions and homology to maize ESTs or blastx alignments to known proteins from the GenBank database. Data of inter-specific comparison of the non shared genes to rice are also shown. "Incomplete" indicates partial hit with the database entry.

B73 locus9002 (AY664413)

Cluster	Gene	Exon #	Start (5')	End (3')	Length	Strand	Intron/Exon boundary
KLM (3085bp)	K	1	196414	196689	276bp	plus	*
	L	1	197389	197688	300bp	plus	*
	M	1	199066	199326	261bp	plus	#
		2	199436	199498	63bp		

54> <55 56> <57 58> <59

[K] [L] [M1] [M2]

Cluster	Gene	Exon #	Start (5')	End (3')	Length	Strand	Intron/Exon boundary
NOPQ (3442bp)	N	1	206783	206625	159bp	minus	*
	O	2	207293	207192	102bp	minus	#
		1	208078	207727	352bp		
	P	2	208730	208577	154bp	minus	#
		1	209052	208844	209bp		
	Q	3	209474	209356	119bp	minus	#
		2	209825	209600	226bp		
		1	210066	209947	120bp		

34> 44> <43 <36 38> <37 71> <39 40><72 <41

[N] [O2] [O1] [P2] [P1] [Q3] [Q2] [Q1]

Cluster	Gene	Exon #	Start (5')	End (3')	Length	Strand	Intron/Exon boundary
RST (6971bp)	R	1	240761	241012	252bp	plus	#
		2	241757	241816	60bp		
	S	1	242769	242912	144bp	plus	#
		2	242984	243133	150bp		
	T	1	244944	245024	81bp	plus	#
		2	245096	245323	228bp		
		3	246246	246314	69bp		
		4	246575	246700	126bp		
		5	247148	247336	189bp		
		6	247429	247731	303bp		

21> <22 <32 25> <24

[R1] [R2] [S1] [S2] [T1] [T2] [T3] [T4] [T5] [T6]

Cluster	Gene	Exon #	Start (5')	End (3')	Length	Strand	Intron/Exon boundary
GHIJ (6728bp)	G	1	180804	180959	156bp	plus	*
	H	1	181527	181733	207bp	plus	*
	I	1	184106	184237	132bp	plus	#
	J	1	187280	187531	252bp	plus	*

48> <49 52> <53

[G] [H] [I] [J]

B73 locus9008 (AY664414)

Cluster	Gene	Exon #	Start (5')	End (3')	Length	Strand	Intron/Exon boundary
HI (4967bp)	H	5	83076	82816	261bp	minus	#
		4	83232	83188	45bp		
		3	83711	83550	162bp		
		2	83869	83801	69bp		
		1	84444	84259	186bp		
	I	4	86236	85865	372bp	minus	#
		3	86585	86475	111bp		
		2	86824	86693	132bp		
		1	87782	87714	69bp		

16> 28> <27 <29 19> <20

[H5] [H4] [H3] [H2] [H1] [I4] [I3] [I2] [I1]

Mo17 locus9009 (AY664419)

Cluster	Gene	Exon #	Start (5')	End (3')	Length	Strand	Intron/Exon boundary
RS (3307bp)	R	3	109587	109330	258bp	minus	#
		2	109741	109673	69bp		
		1	110003	109778	226bp		
	S	5	110245	110130	116bp	minus	#
		4	110580	110495	86bp		
		3	111151	111006	146bp		
		2	112242	112221	22bp		
		1	112636	112345	292bp		

13> 11> <12 14> <15

[R3] [R2] [R1] [S5] [S4] [S3] [S2] [S1]

Legend: * Intron/Exon boundaries unclear; # GT_AG Inton/Exon boundaries detected

Table 2: Annotation and schematic representation of the non-shared gene clusters at loci 9002, 9008 and 9009. The 5'- and 3'-end positions correspond to the nucleotide position in the deposited contig sequences AY664413 (B73-9002), AY664414 (B73-9008), and AY664419 (Mo17-9009). The positions of the specific primers are also reported (not to scale): different genes within each cluster are reported with different colors; blocks and diagonals represent correct exons and introns, respectively.

Genomic PCR

Single non-shared genes are present in both maize inbred lines and occur in multiple copies in the genome

PCR was performed using the non-shared gene specific primer sets designed from the available sequences of the B73 or the Mo17 allele, in high stringency conditions. For all the single genes, the product of expected size was obtained from genomic DNA of both inbred lines B73 and Mo17 (Figure 1 and Table 3). Therefore, very close homologues of the single genic elements non-shared between B73 and Mo17 at the locus sequenced by Brunner (Brunner et al. submitted), appear to be present in both B73 and Mo17 genomes. To estimate the copy number of these amplicons in the maize genome, the same primer sets were used in a PCR screening of a complete set of oat-maize chromosome addition lines (Kynast et al. 2001), and the amplification pattern is summarized in Table 3.

When tested on oat-maize addition lines, the single-gene primer sets showed positive bands on multiple chromosomes (typically from 3 to 6 copies; see Table 3). The genes M9002 (#58-59) and Q9002 (#40-41) are exceptions, mapping only to a single chromosome. The primer set used for gene S9009 (#14-15) produced no positive band in the oat-maize addition lines and thus could not be mapped, even though it amplified correctly on the donor line. The occurrence of a certain number of chromosomal rearrangements and deletions in the maize chromosomes present in the addition lines can not be excluded (Ananiev E.V., personal communication), and might account for the observed pattern for primer set #14-15.

Interestingly, PCR on the oat-maize addition lines using primers specific for the genes shared between the two inbred lines always amplified from only a single maize chromosome, confirming the single copy nature of these genes. This is consistent with the inter-specific comparison data that assigned most of the shared genes within the loci 9002, 9008 and 9009 to the same synthenic regions in rice (Brunner et al. submitted), while for the non-shared genes a break in the maize/rice colinearity was observed (Table 1).

The clusters of non-shared genes are present in both inbred lines genomes at a lower copy number.

Similarly to what was observed for single genes, also all PCR assays using primer sets across neighboring genes within clusters gave positive amplification results on genomic DNA of both maize inbred lines. Exceptions are the combination of primers on

Q9002-exon1 with any other gene belonging to the NOPQ9002 cluster (#44-41, from O9002; #38-41 and #71-41 from P9002) and primers in gene K9002 with gene L9002 (#54-57) within the KLM9002 cluster, where no amplification was observed in Mo17. Therefore both the gene arrangements are specific for B73.

Generally, amplification across adjacent genes which form clusters produced amplification products from a smaller number of chromosomes than amplification using single-gene specific primer sets on genomic DNA of oat-maize addition lines (see Table 3). The whole clusters NOPQ9002 and KLM9002 were mapped on chromosome 1, while the clusters HI9008 and RST9002 mapped on two different chromosomes each (chromosomes 2 and 8, and 2 and 6, respectively). The cluster GHIJ9002 was not assigned to any chromosome due the lack of success in amplifying across the large distance in-between single genes. Further, the cluster RS9009 was also not mapped, since the primer set #11-15 and the #14-15, failed to amplify from any oat-maize chromosome addition lines, even though the amplification from the donor line was positive. The oat-maize addition lines mapping results for clusters KLM9002, NOPQ9002 and one of the two copies of the cluster HI9008 confirmed the chromosomal location of their original locus in B73 (loci 9002 and 9008 on chromosomes 1 and 2, respectively). In contrast, the two copies of the RST9002 cluster were both mapped to chromosomal locations different from the map position of locus 9002 in B73. As expected, each of the mapped clusters was always identified on chromosomes which were also positive for its single genes, while the opposite was never observed. This confirms the reliability of the oat-maize addition lines mapping approach. For further validation, all the amplification products from oat-maize addition lines were also sequenced. Amplicons specific for both single non-shared genes and non-shared clusters showed 95% to 99% sequence similarity when aligned and compared with the corresponding sequences available from the B73 or Mo17 contigs.
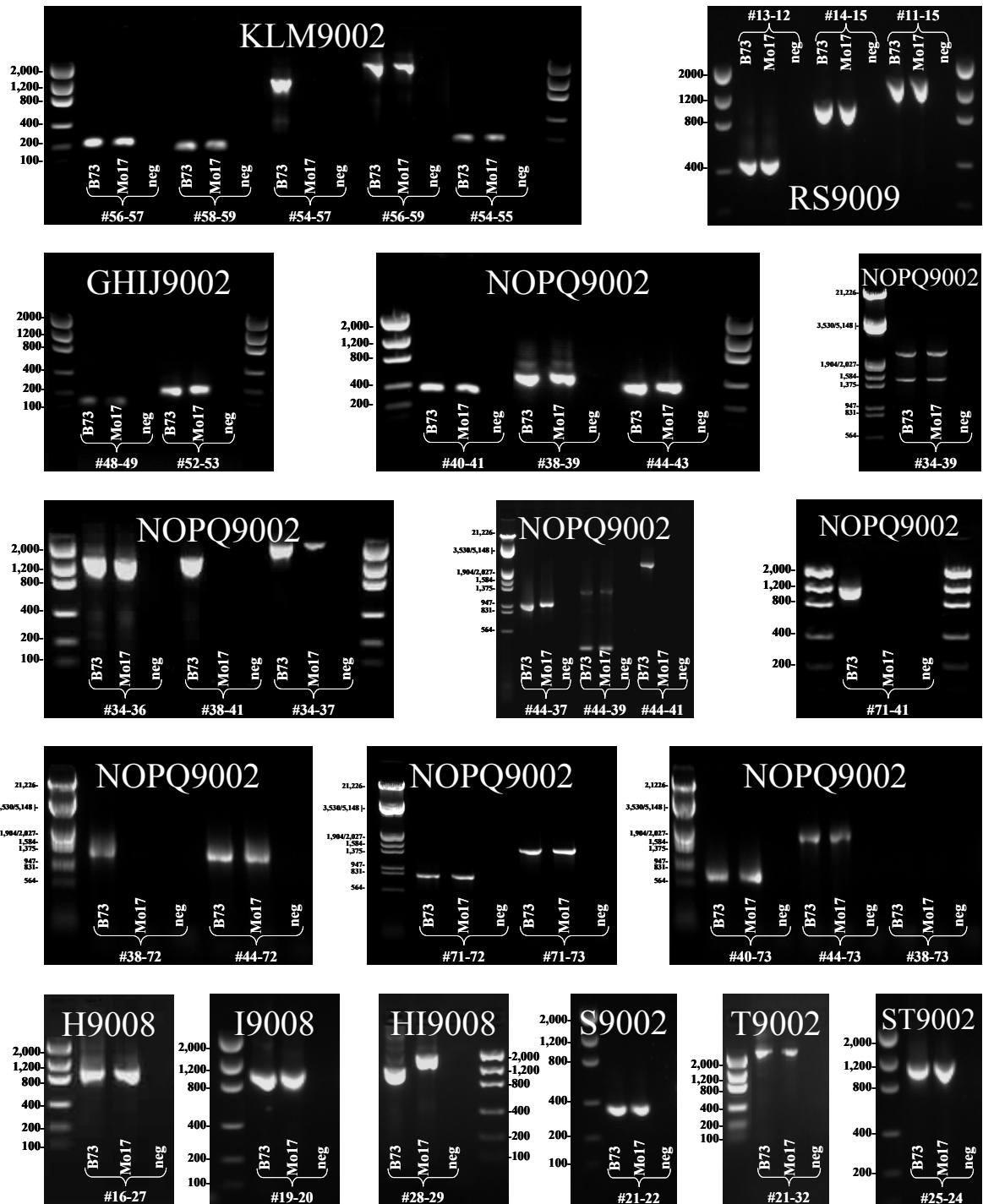
Figure 1: Gel electrophoresis of genomic PCR product on B73 and Mo17 inbred lines (1.5% agarose gel, ethidium bromide staining); neg.: PCR negative control (water).

Gene Cluster	Amplicon	Exp. Size (bp)	Genomic PCR		Oat/maize chromosome addition lines											
			B73	Mo17	maize	oat	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10
KLM (B73-9002)																
K9002	#54-55	204	yes	yes	+		+				+			+		+
L9002	#56-57	211	yes	yes	+		+				+	+				
M9002	#58-59	157	yes	yes	+		+									
K9002 + L9002	#54-57	1174	yes	**no**	+		+									
L9002 + M9002	#56-59	1892	yes	yes	+		+									
NOPQ (B73-9002)																
O9002	#44-43	308	yes	yes	+		+				+		+	+		
P9002	#38-39	427	yes	yes	+		+		+	+	+	+				+
Q9002	#40-41	304	yes	yes	+		+									
N9002 + P9002	#34-37	2051	yes	yes	+		+									
N9002 + P9002	#34-39	2330	2 bands	2 bands	+		+									
N9002 + O9002	#34-36	1374	yes	yes	+		+									
O9002 + P9002	#44-37	957	yes	yes	+		+									
O9002 + P9002	#44-39	1236	2 bands	2 bands	+		+									
O9002 + Q9002	#44-41	2288	yes	**no**	+		+									
P9002 + Q9002	#38-41	1479	yes	**no**	+		+									
P9002 + Q9002	#71-41	1081	yes	**no**	+		+									
P9002 + Q9002	#38-72	1211	yes	**no**	+		+									
O9002 + Q9002	#44-72	2020	yes	yes	+		+									
P9002 + Q9002	#71-72	813	yes	yes	+		+									
GHIJ (B73-9002)																
H9002	#48-49	109	yes	yes	+		+				+		+	+		
J9002	#52-53	201	yes	yes	+		+	+	+			+	+			+
RST (B73-9002)																
S9002	#21-22	344	yes	yes	+	+		+				+		+		
T9002	#25-24	992	yes	yes	+			+	+			+				
S9002 + T9002	#21-32	2194	yes	yes	+			+				+				
HI (B73-9008)																
H9008	#16-27	842	yes	yes	+			+				+	+	+	+	
I9008	#19-20	897	yes	yes	+	+		+						+		
H9008 + I9008	#28-29	1678	yes	yes	+			+						+		
RS (Mo17-9009)																
R9009	#13-12	433	yes	yes	+				+	+	+			+		+
S9009	#14-15	960	yes	yes	+											
R9009 + S9009	#11-15	1339	yes	yes	+											

Table 3: Summary of PCR results on genomic DNA and oat-maize chromosome addition lines. + indicates positive amplification on the correspondent addition line; primer sets across neighboring genes within a cluster are underlined; frames around positives indicate chromosomes carrying the specific non-shared gene clusters. Some primer sets produced amplification products from all the oat-maize chromosome addition lines, most probably corresponding to repetitive sequences amplification. These primer sets are not reported in the table.

The NOPQ-9002 cluster is present in two different copies in both inbred lines

Interestingly, the PCR amplification of primer sets #34-39 and #44-39, which comprise the exon P2 of gene P9002, produced a double band pattern in both inbred lines (2,330bp and 1,236bp respectively, Figure 1). The size of the higher band was consistent with the expectation for locus 9002 in B73, while the lower band was in both cases ca. 850bp shorter. This observation suggested the presence of a second locus in the genome of both inbred lines, but carrying a deletion somewhere between the primer #44 (exon O1) and the primer #39 (exon P1). Further, the fact that primer set #34-37 and #44-37 produced only a single band of the expected size in both lines, suggested that the deletion in the second locus might involve the region including the primer #37 (exon P2).

The genomic PCR products #44-39 (both bands) and #34-36 from B73 were cloned and sequenced. The sequence of the 1,236bp (higher) band from #44-39 was identical to the corresponding B73 locus 9002 and the sequences of the #34-36 clones identified two haplotypes, differing by number of single nucleotide polymorphisms (SNP) and INDELs. The 280bp overlap between the sequences of #34-36 and #44-39 amplicons (see Table 2 and Table 5), as well as the presence of locus-specific SNPs and INDELS, allowed the assembly of the #44-39 and #34-36 amplicons into two contigs of 2,330bp and 1,494bp, which correspond to the sequences between primers #34 and #39 at two loci. The first sequence is that of the locus 9002 from B73, while the second one might represent the second locus of this non-shared gene cluster. When aligned to the B73 9002 locus, the latter sequence showed a 16bp insertion substituting the 855bp region from position 208,064 to 208,918 of 9002. The corresponding deleted region of the B73 locus 9002 stretches from the last 15bp of exon O1 (gene O9002) to the first 75bp of exon P1 (gene P9002), and includes the whole exon P2 of gene P9002. Consequently, in this second locus, part of the exons O1 (gene O9002) and part of the exon P1 (gene P9002) are fused. Similar results were obtained from the sequencing of the corresponding genomic PCR products from Mo17, with the exception of some SNPs and INDELs, confirming the presence of the two different gene clusters also in this inbred line. A blastn analysis on the sequence of the second locus of the #34-39 B73 assembly against the TIGR Maize Genome Project Database, produced a significant alignment to the AZM4_123277 from B73 (total length: 2,431bp; identities: 917/918, 99%; *plus/minus* orientation from position 577 to 1,494 on the #34-39 assembly, corresponding to position 2,431 to 1,514 on the AZM4_123277 sequence respectively). When aligned to the B73-9002 contig sequence, the alignment with the AZM4_123277 extended from position 207,252 (corresponding to

position 577 on the #34-39 assembly and 2,431 of the AZM4_123277 sequence) to 208,063 (corresponding to the left border of the observed deletion between the B73-9002 and the second locus cluster), and from position 208,919 (corresponding to the right border of the deletion) to 209,898 (299bp downstream exon Q2 of gene Q9002, and corresponding to position 622 on the AZM4_123277 sequence). The borders of the two alignments between the AZM4_123277 sequence and the B73-9002 contig identified a 16bp sequence which corresponded to the short insertion found in the second locus. The 16bp insertion is present in both B73 and Mo17 copy of the second locus. A large part of the 5' end of AZM4_123277 sequence extends beyond the #34-39 B73 assembly. However, genomic PCR reactions performed with a specific primer designed to this region (#73) and the other primers specific for the cluster (#40, #71, #44, and #38, see Figure 1) confirmed that the AZM4_123277 might correspond in both inbred lines to regions presenting this second locus. Preliminary data (not shown) indicate that also this copy of the cluster, like the 9002 copy, may map in different genomic positions in B73 and Mo17 (Figure 2). The PCR results #38-41 and #38-72 indicate that the gene Q9002 is missing from the cluster in the locus carrying the exon P2 (gene P9002) in Mo17.

A schematic representation for both the clusters in Mo17 and B73, as inferred from genomic PCR results (presence-absence of bands and band size comparison), is summarized in Figure 2.
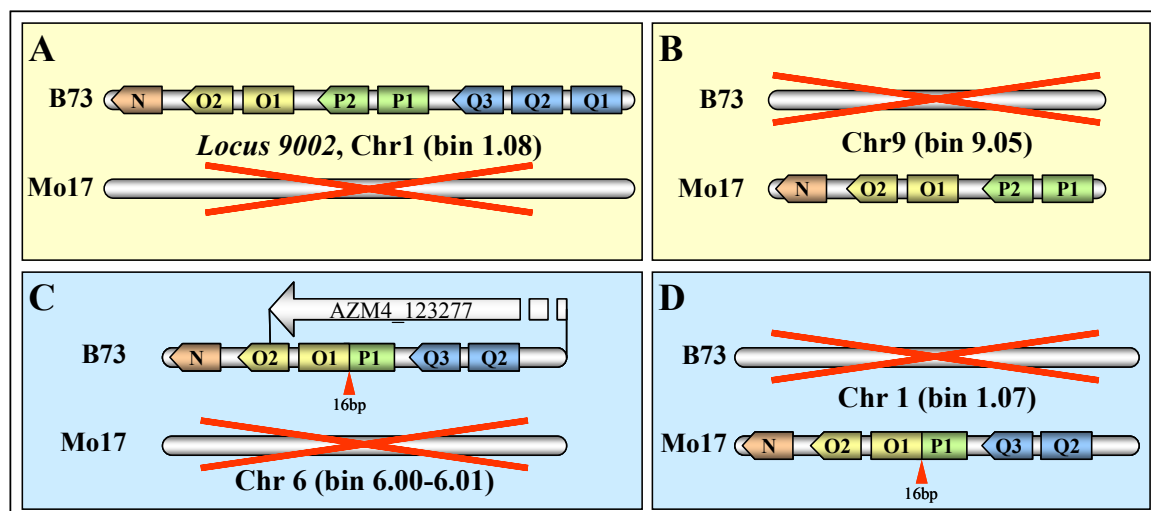
Figure 2: Schematic representation of the copies of the cluster NOPQ present in B73 and Mo17. **A, B**: Structures of the *Locus 9002*-like copies. **C, D**: Structures of the second copy of the NOPQ cluster; exons O1 and P1 are fused due to the 855bp deletion including Exon P2, the Intron between P1 and P2 and the region between P2 and O1 including the terminal portions of exon O1 and P1; the red triangle indicates the 16bp insertion at exon O1-P1 junction (see text); alignment of the B73 copy with the B73 AZM4_123277 is indicated with a grey arrow; preliminary data (not shown) indicate that also this copy is not collinear between the two inbreds. Positions of the four loci as inferred from both physical and genetic mapping are reported.

Expression analysis

The expression status of the non-shared genes was analyzed in an RT-PCR on mRNA from different tissues (seedlings, roots and leaves) collected from B73, Mo17 and their reciprocal hybrids. Expression pattern from reciprocal hybrids were used as indicator of possible epigenetic effects on the gene expression. Two independent biological replicates for each tissue and each genotype were analyzed. The same primer sets already described above were used. The RT-PCR bands are visualized in Figure 3, Figure 4 and Figure 5. The results of expression analysis are summarized in Table 4.
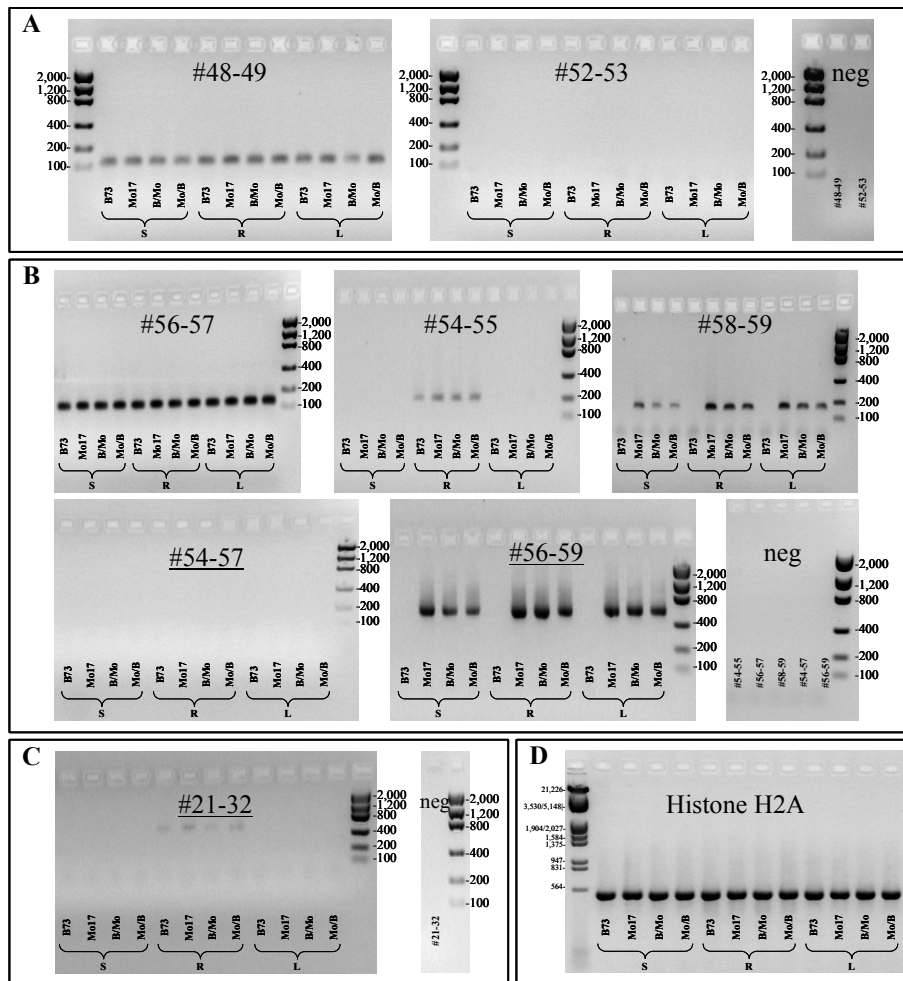
Figure 3: Gel electrophoresis of RT-PCR products (2% agarose gel, ethidium bromide staining, negative image). **A:** cluster GHIJ, locus B73 9002; **B:** cluster KLM, locus B73 9009; **C:** cluster RST, locus B73 9002; **D:** Histone H2A (RT-PCR positive control); neg: RT-PCR negative control (water). Underlined amplicons indicate primer sets designed across genes.
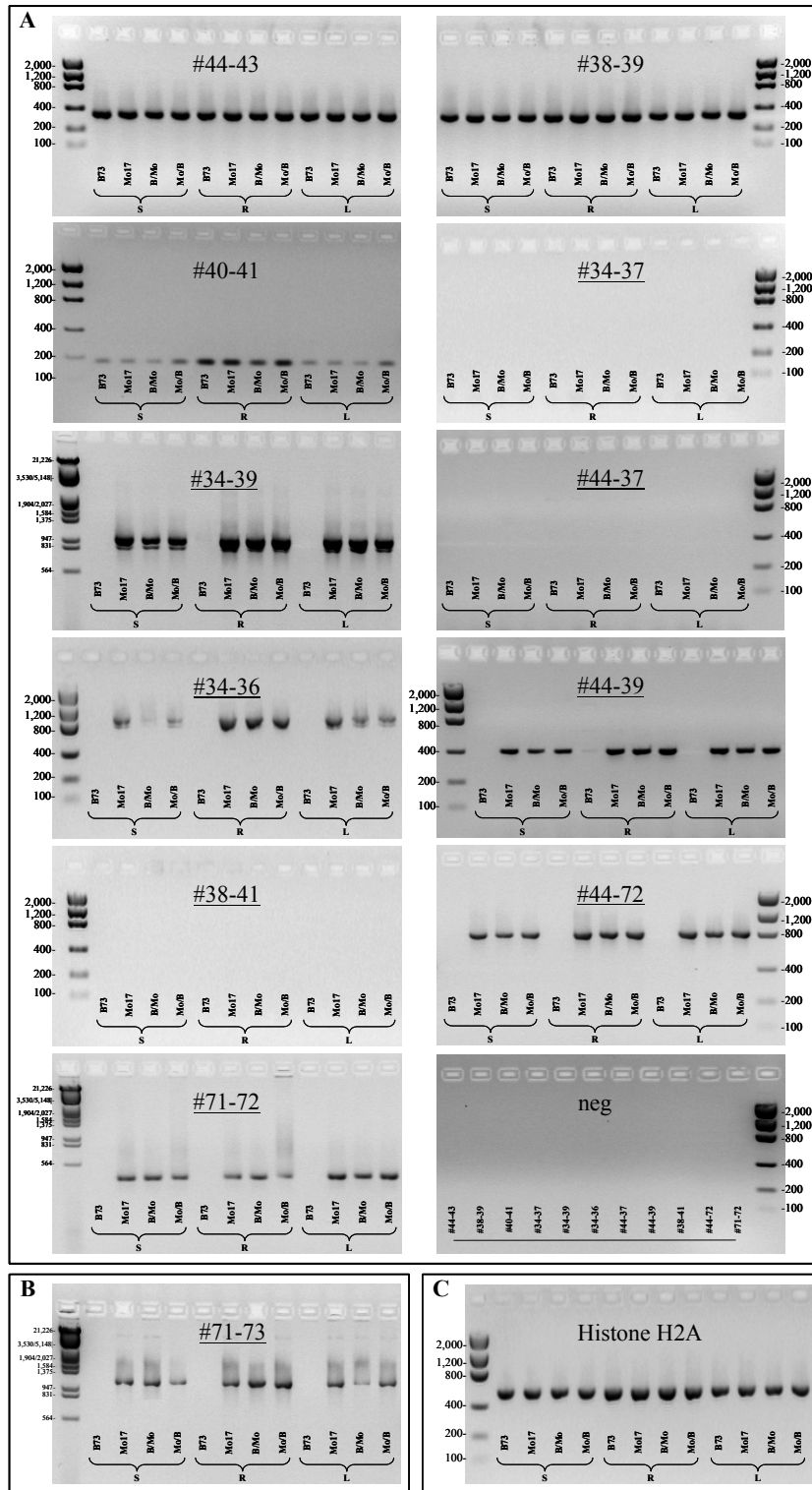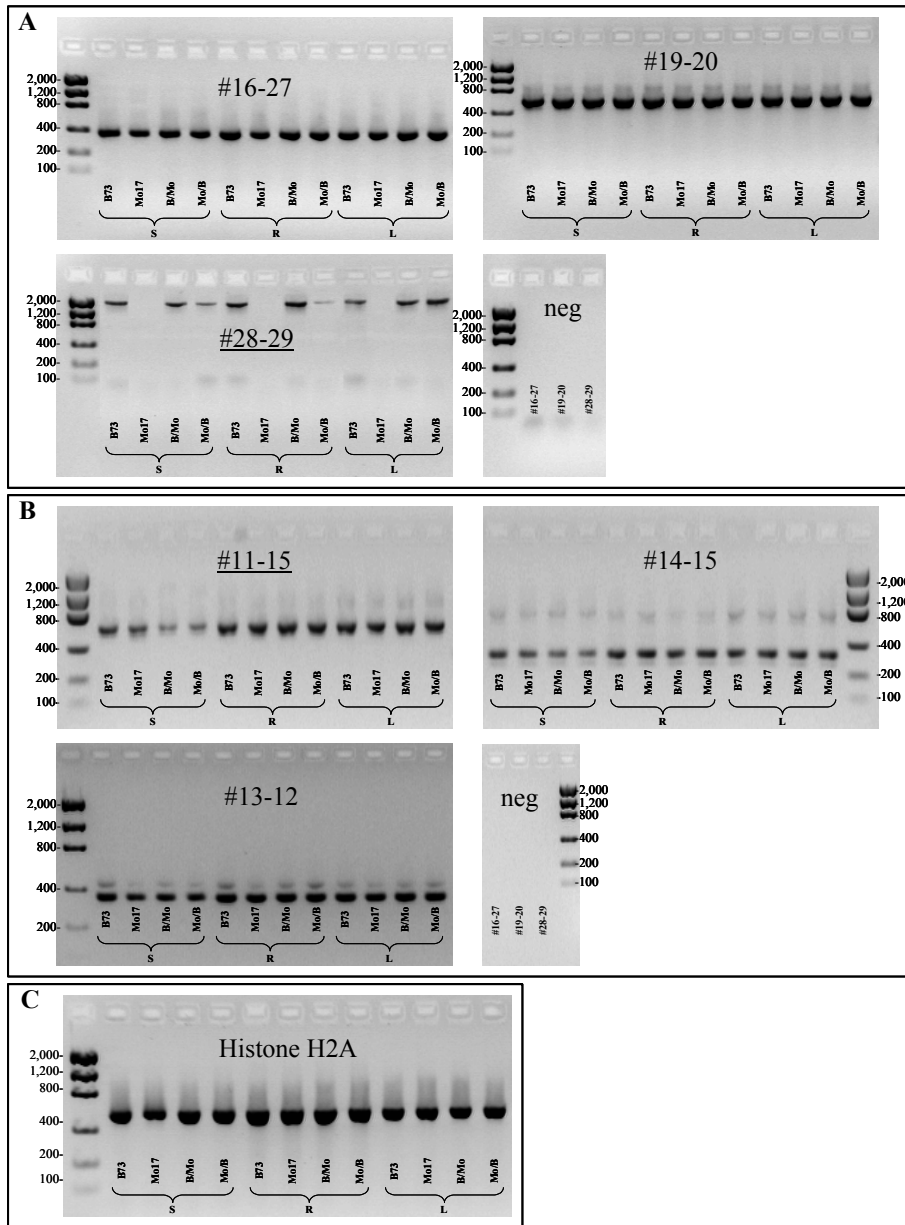
Figure 4: Gel electrophoresis of RT-PCR products (2% agarose gel, ethidium bromide staining, negative image). **A:** cluster NOPQ, locus B73 9002; **B:** specific amplification on AZM4_123277, cluster NOPQ second copy (see text); **C:** Histone H2A (RT-PCR positive control); neg: RT-PCR negative control (water). Underlined amplicons indicate primer sets designed across genes.

Figure 5: Gel electrophoresis of RT-PCR products (2% agarose gel, ethidium bromide staining, negative image). **A:** cluster HI, locus B73 9008; **B:** cluster RS, locus Mo17 9009; **C:** Histone H2A (RT-PCR positive control); neg: RT-PCR negative control (water). Underlined amplicons indicate primer sets designed across genes. Underlined amplicons indicate primer sets designed across genes.

Gene Cluster	Amplicon	Genomic PCR		RT-PCR Seedling (S)				RT-PCR Root (R)				RT-PCR Leaf (L)			
		B73	Mo17	B73	Mo17	B73Mo17	Mo17/B73	B73	Mo17	B73Mo17	Mo17/B73	B73	Mo17	B73Mo17	Mo17/B73
KLM (B73-9002)															
K9002	#54-55	yes	yes	-	-	-	-	+	+	+	+	-	-	-	-
L9002	#56-57	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
M9002	#58-59	yes	yes	-	+	+	+	-	+	+	+	-	+	+	+
K9002 + L9002	#54-57	yes	**no**	-	-	-	-	-	-	-	-	-	-	-	-
L9002 + M9002	#56-59	yes	yes	-	+	+	+	-	+	+	+	-	+	+	+
NOPQ (B73-9002)															
O9002	#44-43	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
P9002	#38-39	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
Q9002	#40-41	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
N9002 + P9002	#34-37	yes	yes	-	-	-	-	-	-	-	-	-	-	-	-
N9002 + P9002	#34-39	2 bands	2 bands	-	+	+	+	-	+	+	+	-	+	+	+
N9002 + O9002	#34-36	yes	yes	-	+	+	+	-	+	+	+	-	+	+	+
O9002 + P9002	#44-37	yes	yes	-	-	-	-	-	-	-	-	-	-	-	-
O9002 + P9002	#44-39	2 bands	2 bands	-	+	+	+	-	+	+	+	-	+	+	+
O9002 + Q9002	#44-41*	yes	**no**	-	-	-	-	-	-	-	-	-	-	-	-
P9002 + Q9002	#38-41	yes	**no**	-	-	-	-	-	-	-	-	-	-	-	-
P9002 + Q9002	#71-41*	yes	**no**	-	-	-	-	-	-	-	-	-	-	-	-
P9002 + Q9002	#38-72*	yes	**no**	-	-	-	-	-	-	-	-	-	-	-	-
O9002 + Q9002	#44-72	yes	yes	-	+	+	+	-	+	+	+	-	+	+	+
P9002 + Q9002	#71-72	yes	yes	-	+	+	+	-	+	+	+	-	+	+	+
GHIJ (B73-9002)															
H9002	#48-49	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
J9002	#52-53	yes	yes	-	-	-	-	-	-	-	-	-	-	-	-
RST (B73-9002)															
S9002	#21-22	yes	yes	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.	n.r.
S9002 + T9002	#21-32	yes	yes	-	-	-	-	+	+	+	+	-	-	-	-
HI (B73-9008)															
H9008	#16-27	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
I9008	#19-20	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
H9008 + I9008	#28-29	yes	yes	+	-	+	+	+	-	+	+	+	-	+	+
RS (Mo17-9009)															
R9009	#13-12	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
S9009	#14-15	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+
R9009 + S9009	#11-15	yes	yes	+	+	+	+	+	+	+	+	+	+	+	+

* lack of expression have been inferred from concurring evidences coming from the other primer sets in the cluster.
n.r.: not reproducible

Table 4: Summary of RT-PCR results using primer sets within and across neighboring non-shared genes (underlined). The corresponding genomic amplification is noted in columns 3 and 4.

The single non-shared genes are generally expressed in both maize inbred lines

The primer sets specific for non-shared genes have amplified a product in all RT-PCR reactions from both inbred lines (Figure 3, Figure 4, Figure 5, and Table 4). In some cases, the smaller size of the RT-PCR bands compared to the genomic PCR product size (Figure 1, Figure 3, Figure 4 and Figure 5) suggests that splicing of the relative mRNA has occurred. Further, multiple RT-PCR bands were detected for some of the non-shared genes, which might either originate from alternative splicing or from the transcription of different copies of the non-shared gene.

Transcription across clustered non-shared genes occurs and differences in expression between the two inbred lines are observed.

RT-PCR products were detected for some of the primer sets, which amplify across clustered non-shared genes (Figure 3, Figure 4, Figure 5, and Table 4). Multiple band patterns were observed for combinations #34-36 and #34-39. Interestingly some of the gene combinations are differentially expressed between the two inbred lines. In particular, the combination #56-59 (from L9002 to M9002) within the cluster KLM9002 (B73 locus 9002), and the combinations #34-39 (from N9002 to P9002 exon-1), #34-36 (from N9002 to O9002) and #44-39 (form O9002 to P9002 exon-1), within the cluster NOPQ9002 (B73 locus 9002) are only expressed in Mo17. Conversely, the primer set #28-29, designed across the B73 locus 9008 specific genes H9008 and I9008, gives positive bands in all the tissues in B73 and in both hybrids, but never in Mo17. The observed expression patterns suggest that, despite the presence of these non-shared gene clusters in one (B73), but not the other of the sequenced alleles, suggests that in general the additional homologous copies of the gene clusters as well as those present at loci 9002, 9008 and 9009 might be expressed.

If differences in expression were detected between the inbred lines, the expression always resulted dominant, i.e. positive bands were detected also in the hybrids when present in at least one of the inbred lines, suggesting absence of trans or epigenetic effects on the expression between the alleles of the tested genes.

Only one of the NOPQ-9002 copies is expressed in Mo17

As mentioned above, differential expression between the two inbred lines was observed consistently for most of the primer sets across non-shared genes at the NOPQ9002 cluster. Further, the polymorphisms between the two genomic copies reported above allowed to assign RT-PCR products to specific genomic copies. No RT-PCR product was amplified

from either B73 nor Mo17 using any primer of the NOPQ cluster in combination with primer #38 or #37, located in geneP9002-exon2 (#34-37, #38-41 and #44-37, Figure4A). All these primer sets are specific for the NOPQ cluster of B73 (locus 9002) and Mo17, which includes P9002-exon2 and/or gene Q9002-exon1 (see scheme in Figure 2). Therefore these loci are not providing the transcripts detected in RT-PCR (i.e. none of the loci amplified from genomic DNA by this primer sets was expressed either in B73 or Mo17 in the tested samples). In addition, the consistent absence of positive RT-PCR band in B73 for all the primer combinations designed across different genes within the cluster and for the primer set #71-73 specific for the AZM4_123277 sequence (Figure 4B) led to the conclusion that neither of the two copies of the cluster was expressed in B73 in the tested conditions. In contrast, positive RT-PCR products were obtained only from inbred Mo17, when using primer combinations that are either specific for the second copy of the NOPQ cluster (based on AZM4_123277 sequence: #71-72 and #44-72) or do not discriminate between the NOPQ cluster copies (#34-39, #34-36, #44-39, Figure 4A). The #44-39 and #34-36 RT-PCR products from Mo17 were sequenced and the partial overlapping sequences were assembled, resulting in two #34-39 contigs of 862bp and 963bp in length, respectively. Based on the comparison of the RT-PCR sequences with the Mo17 #34-39 genomic sequence corresponding to the locus bearing the deletion of gene P9002-exon2, this difference could be assigned to an alternatively spliced intron of 100bp within the second exon of the 963bp transcript and confirmed that it represents the expressed locus in Mo17. Interestingly, the #34-39 RT-PCR sequences were found partially homologous to a rice expressed protein with alternative splicing (gi:50919207). A scheme of the #34-39 alternative transcripts and their alignment with the Mo17 genomic sequence, as well as the comparison of the observed exon-intron structure versus that predicted for the cluster B73-NOPQ9002, are reported in Figure 6.
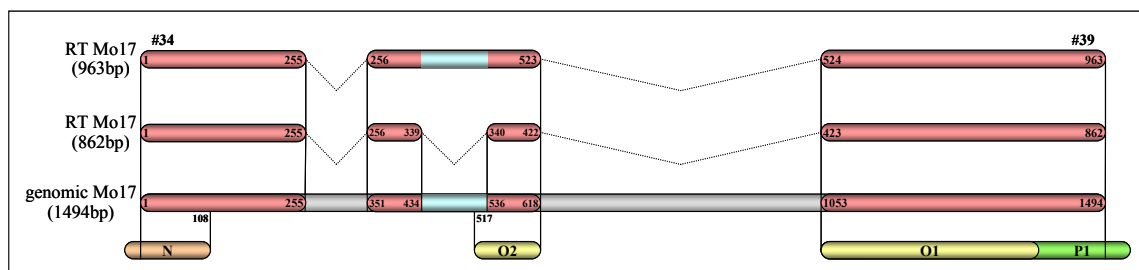

Figure 6: Schematic representation of the alignment between the 862bp and the 963bp RT-PCR contigs #34-39 from Mo17 seedlings and the corresponding genomic sequence. Red blocks represent the observed exons structure within the analyzed sequences. The alternatively spliced intron is represented in blue. The 5'-3' orientation of the genomic sequence is maintained (direction of transcription from right to left). The position of the exons of genes N9002, O9002 and P9002 as preliminary predicted by blastx analysis on the B73 9002 contig sequence is also shown.

Discussion

Structure and origin of non-shared genes

Since several intraspecific sequence comparisons have discovered genic and non-genic allelic non-homologies among maize inbred lines (Fu and Dooner 2002; Song and Messing 2003; Brunner et al. submitted), detailed analysis of their structure, origin and expression is needed for further understanding of this phenomenon in the maize genome. Here, we focus on clustered non-shared genes earlier identified at loci 9002, 9008 and 9009 between maize inbreds B73 and Mo17 (Brunner et al. submitted). Based on the homologies of these non-allelic genes to proteins from the GenBank database, maize ESTs and conserved Exon/Intron boundaries, the possibility that these genic sequences are simply retroposed duplicates of ancestor genes can be excluded. This is in contrast to findings in human, yeast and Drosophila (Long et al. 2003; Wang et al. 2004), where retroposition accounts for large numbers of processed copies of gene duplicates or pseudogenes and it is believed to be one of the driving forces for gene evolution (Brosius 1991). In general, our data indicate that non-shared genes in maize seem to be incomplete, but in many cases more than one intact exon is present (Fu and Dooner 2002; Ramakrishna et al. 2002). The non-shared genes violate the allelic organization of gene order at the analyzed loci. Interestingly, positive PCR amplifications with non-shared gene specific primer sets on both inbred lines, as well as on maize-oat addition lines, confirms that partial or full-length copies of the non-shared genes must exist at additional location across the maize genome. Therefore, non-shared genes might be members of gene families. Similar behavior have already been reported for at least two of the four genes in the non-shared gene cluster at the *bz1* locus (Fu et al. 2001) and for six of the 13 incomplete clustered genes at the Rp1 in maize (Ramakrishna et al. 2002).

The utility of maize-oat addition line PCR data for the assignment of chromosome positions of non-shared genes and their homologues is somehow limited, since the donor line Seneca may represent a haplotype different from either B73 or Mo17. However, these data may serve as an estimate of the copy number to be expected for non-shared genes in the maize genome.

All the analyzed non-shared genes interrupt maize-rice collinearity, as has already been reported (Brunner et al. submitted), while clustering of the rice homologues of non-shared genes was not observed. Since rice is believed to represent the ancestral condition

these observations suggests that the non-shared genes might have originated from insertion events in maize after its divergence from rice.

Even more surprising is that sequences homologous to and collinear with four of the six analyzed clusters of non-shared genes are present elsewhere in the genome of the inbred, which lacks the allele at loci analyzed by Brunner et al. Multiple copies of most clusters exist based on the maize-oat addition line PCR data, even if in a lower copy number compared to the single non-shared genes. Such a structure and organization of non-shared genes has not yet been described for any plant species and appears to be unique to maize. These clusters might either be the result of independent duplications or emerged via a novel rearrangement mechanism, which in maize yields many different inbred haplotypes. Although such a mechanism is not known yet, its effects could have tremendous influence on the sequence variety among modern maize inbreds.

Expression of non-shared genes

If functionally transcribed, the non-shared genes may have a strong phenotypic impact on inbred lines and their hybrids. Using RT-PCR, we found that the majority of the single non-shared genes we analyzed (9 out of 12, 75%) are expressed in all the tested samples (92% including genes that show tissue or genotype specific expression). However no evidence for the specific expression of the non-shared genes in the analyzed clusters could have been produced using RT-PCR, due to the presence of multiple copies in the genome, sharing high level of sequence similarity. Analogous result where previously observed for the non-shared genes between B73 and McC in the *bz* locus (Fu et al. 2001): for at least two of the 4 non-shared genes, multiple cDNA sequences were detected, confirming the expression of several different members of large gene families.

The clustered and directional arrangements of the non-shared genes in the loci we analyzed, as well as their Exon/Intron-like structures, suggested that they might be transcribed as single mRNA. In this case, they might represent an example of novel gene products *in statu nascendi*, originating through the shuffling and fusion of different pre-existent protein domains (Blake 1983). Interestingly, RT-PCR products were observed for all the clusters for which primer sets could be designed across exons of different non-shared genes, indicating that they might be transcribed as single mRNA. The consistent RT-PCR patterns observed for the primer sets designed across different genes within the cluster NOPQ locus 9002 B73 also support this hypothesis. The analysis of the partial sequence available for the two copies of the cluster NOPQ allowed us to identify which

copy is expressed in Mo17, and to define the real exon/intron junction structure of the "chimeric" mRNA, whose correct splicing was thus confirmed. While we do not know if the copy of the NOPQ cluster expressed in Mo17 encodes a functional polypeptide, we found not evidence of nonsense or frame shift mutations inconsistent with functional expression. Further evidence of the presence of full-length cDNA as well as from the analysis of the relative protein levels must be produced, even though it is possible that an untranslated transcript may still play a regulatory role. Finally, results confirm that both the specific non-shared copy within the locus 9002 B73 and its counterpart in Mo17 are not expressed in the tested samples. The absence in the public databases of expressed sequences overlapping the single genes of cluster NOPQ locus 9002 B73, for which they were annotated as individual genes, is probably due to the fact that most of the public sequences are derived from B73, while the transcription of the cluster is only detectable in Mo17, at least for the tested samples.

The arrangement of some maize genes as allele-specific clusters in different locations within the maize genome, and preliminary PCR evidence suggest that many or all of the copies might be non-collinear between maize inbred lines. The transcription of each copy of the clusters may be greatly influenced by differences in the surrounding genetic environments, which may lead to gain or lost of specificity and functionality (Langham et al. 2004). In fact, RT-PCR across the non-shared clusters shows difference in the expression between the two inbred lines for 3 of the 5 analyzed clusters. Noticeably, lack of expression has been observed either in B73 or in Mo17, irrespective of which inbred contained the gene in the locus analyzed by Brunner et al. The different expression pattern between the two copies of the cluster NOPQ within Mo17, as well as between B73 and Mo17, is particularly evident. This suggests a possible correlation between non-collinearity and differences between transcriptional patterns, which thus might be a common feature of gene clusters corresponding to non-shared genes. Promoters of gene-adjacent LTR-retrotransposons, which represent most of the intergenic variability in maize inbreds, have been suggested to be responsible for a distinct expression profile (Llave et al. 2002; Dunn et al. 2003; Schramke and Allshire 2003). Different gene-adjacent sequences may exert a different *cis*-acting regulatory influence, for example causing tissue specificity or temporal regulation of expression (Birchler et al., 2003; Dunn et al., 2003; Schramke and Allshire, 2003; Kazazian, 2004; Knight, 2004). The intranuclear position of many genes has also been correlated with their activity state, suggesting that it may influence gene expression (Osborne et al. 2004). The non-shared genes, being in completely different location in the

genome, are more likely to be affected by different genetic environments than the shared genes, especially in maize. For the analyzed clusters, dominance of expression, observed whenever a difference in expression was detected, indicates that *cis*-acting elements might play the major role. This is in agreement with the observations on allelic preferential expression due to *cis*-effects recently observed in maize hybrids (Guo et al. 2004). Non-shared genes might be considered a special case of allelic regulation, for which *cis*-effects are particularly enhanced by non-allelism. However, epigenetic or *trans* effects cannot be generally ruled out as factors possibly intervening in the general transcription regulation of the non-shared genes in the inbred lines as well as in their hybrids.

The maintenance of non-expressed copies of the non-shared clusters in both the inbreds suggests that they might probably have a function in tissues, developmental stages or conditions different from the tested ones. Alternatively, a peculiar capacity of maintaining additional non-functional copies of genes in the genome must be invoked for maize; in this case, the usual pattern involving gene duplication, selection and maintenance only of the functional copies (Prince and Pickett 2002) would not be as stringent in maize as it is in other species. This might be justified from an adaptive or evolutionary point of view, considering the strong artificial selection process during maize domestication and breeding.

Non-collinearity and hybrid vigor

Mechanisms involving expression regulation have been proposed to be the cause of heterotic complementation (Birchler et al., 2003; Song and Messing, 2003), as well as of the allelic interactions proposed by the overdominance theory for explaining hybrid vigor (Crow, 1948; Song and Messing, 2003). Our findings support the possibility that non-shared genes and intergenic elements may contribute different transcriptomes in different maize inbred lines, in terms of quality, quantity and type of expressed genes. Hybrids resulting from the cross of such lines not only would inherit these unique features, but also develop new regulatory properties resulting from both the fusion and the complementation of the single inbreds characteristics. Heterotic groups differing in their complement of functional genes would be particularly effective when combined in a hybrid. However, this interaction would not simply involve a dominant complementation in the hybrid, implied by allelism, but would most likely consist of a complex integration in the genetic regulation pathways via multiple levels of interaction.

Modern maize inbreds have originated from human selection for the production of the best hybrids. It has been proposed that this process might have led to the selection of loci maximizing heterosis (Duvick 2001). This process might have unconsciously favored the maintenance of high levels of intraspecific diversity among inbreds, including the non-shared genic and intergenic elements. The whole genome and transcriptome would then be affected by non-collinearity, each locus contributing (and/or being affected) differently in different lines, crosses, environment, developmental processes and so forth. In other words, the system of creating, maintaining and bearing non-collinearity might have been subject to selection in maize, i.e. the general molecular mechanism determining heterosis, and not the particular loci, would have been selected. This hypothesis may explain why neither specific loci nor genetic mechanisms determining hybrid vigor have been clearly detected yet. If this is correct, heterosis would be controlled at the genomic rather than at the genic level, and classical theories of heterosis would have to be reconsidered.

Due to the extensive breeding programs, the performance of maize inbreds increased dramatically during the last century. However, heterosis remained the same or only slightly increased. Better inbreds produced better hybrids in absolute terms, but the relative heterosis (how much hybrid is better than the parental lines) was maintained (Duvick 2001). If non-collinearity of both genic and inter-genic regions had a role in determining heterosis, the improved inbreds should have been simultaneously selected for the maintenance of good cross-mating capabilities, and for the expansion of non-collinearity among lines which contributed to the high performance. We suggest here a balancing selection between reduction of non-collinearity (= less heterosis) vs. too much non-collinearity (= lower cross-mating viability).

It is to be noted that this mechanism can function in a heterozygous environment only; and it cannot be fixed by breeding in any single inbred line. Finally, since we propose that heterosis arises from the combinatorial interaction of complex genomic properties, our hypothesis explains why the level of heterosis can not be predicted by the performance of single inbred lines.

Materials and Methods

Plant Materials

Seedlings, juvenile roots and leaves from the two maize inbred lines B73 (Iowa Stiff Stalk Synthetic Population) and Mo17 (Lancaster Sure Crop) and their reciprocal hybrids were collected for RNA extraction (seeds kindly made available by M. Guo, Pioneer Hi-Bred International., Johnston, IA). Material from single individuals was stored in separate tubes, immediately frozen in liquid nitrogen and stored at -80°C. 50 seeds for each genotype were germinated in Agriperlite in 4cm diameter well arrays in growth chambers (26°C constant temperature; photoperiod 14/10 hrs. light/dark; R.H. 90%). Juvenile tissues (*seedlings* and *roots*) were collected from about 35 plants at the same developmental stage (at the appearance of the tip of the 4th leaf). From each individual, samples from the whole whorl above the 1st leaf collar, with the exclusion of the 2nd leaf (referred as *seedlings*, *S*) and radicles and lateral seminal roots (referred as *roots*, *R*) were collected. Roots were rinsed in water and quick dried on absorbent paper towels before storage. The remaining plants were transferred in 20cm diameter pots in common soil and growth at the same conditions until the appearance of the tip of the 6th leaf from the whorl, when the entire 5th leaf was collected (samples referred as *leaves*, *L*).

Gene annotation

The non-shared genes herein analyzed were predicted in sequenced contig 9002, 9008 and 9009 as reported by Brunner et al. (submitted); they were named "non-shared" due to their absence in one or the other of the two compared collinear contigs between the maize inbred lines B73 and Mo17. For the purpose of this work, the exon-intron structures of the non-shared genes were refined aligning the predicted genes sequences against the GenBank database using the translating BLAST (blastx) alignment algorithm (http://www.ncbi.nih.gov/BLAST/) as well as checking for the presence of the GT_AG splicing sites in the available genomic sequence. Discontinuous alignments with the same entry in the database were considered as consecutive exons of the same gene; otherwise they were annotated as distinct genes.

PCR primers design

Primer for PCR assays were designed to the contig sequences using the Primer3 software (Rozen and Skaletsky 2000). Default parameters were used, except for Primer

size 21bp (min 20bp, max 22bp), GC% 58 (min 53%, max 80%), and GC Clamp 1. In order to be feasible for RT-PCR assays, whenever possible, primer sets were designed within regions corresponding to exons, predicted as above. The primers, combined in different sets, were used to perform amplifications within single genes as well as between neighboring genes, whenever the distance between them allowed a reasonable PCR product size (<2.5kb). Tested primer features are shown in Table 5.

Name	dir	len	tm	gc%	seq (5' to 3')	Start pos	Gene	Allele
#11	fwd	20	62,35	50,00	TCGTGAGCCAAACAAACAGC	109799	R9009	Mo17
#12	rev	21	58,98	52,38	TCATGTCAGACTCACCTCTGC	109918	R9009	Mo17
#13	fwd	20	61,69	55,00	CTCCACACTTTTGCGACCAC	109486	R9009	Mo17
#14	fwd	20	60,63	50,00	GCCTTTATTTCCCCATCCAG	110178	S9009	Mo17
#15	rev	21	60,83	52,38	CGTTACCTCGACTTGCATGAG	111137	S9009	Mo17
#16	fwd	21	60,07	52,38	CTGAGGCGAACTTTGTACAGC	83554	H9008	B73
#17	rev	20	60,00	50,00	TAGCTTGGCGTGCATCATAG	84097	H9008	B73
#18	rev	20	60,02	55,00	GGCGTGGAAAAGTAGAGCTG	84440	H9008	B73
#19	fwd	20	59,93	55,00	CTTGTTGTGAGCAGCAGAGC	85922	I9008	B73
#20	rev	20	61,11	55,00	GAGAAGAAGCCCCGTGAATC	86818	I9008	B73
#21	fwd	22	61,80	50,00	GAGCCTTACAATGTCGGTGATG	242790	S9002	B73
#22	rev	20	59,44	50,00	CTAGGCGTTGATCATTGCAG	243133	S9002	B73
#23	fwd	20	59,90	50,00	AGCTTCCAGCTTTTGCTCTG	247221	T9002	B73
#24	rev	21	59,60	52,38	CACGGGATCAGGACTAACTTG	247697	T9002	B73
#25	fwd	21	61,02	57,14	GACCCTTCCCATAGCTCACTG	246706	T9002	B73
#26	fwd	21	59,93	52,38	GTACCTGTCAAAGACGCTTGC	83846	H9008	B73
#27	rev	22	59,41	50,00	GCCAAGTTACCTTCTCCAAGAG	84395	H9008	B73
#28	fwd	21	58,81	52,38	TACTCGTACATCCTGCCACTG	84275	H9008	B73
#29	rev	20	60,01	55,00	TCAGTATGGTGGCTCTGCTG	85952	I9008	B73
#30	fwd	22	58,28	50,00	CCCTCAGCTAGAAAAGTCAGTG	86903	I9008	B73
#31	rev	21	58,85	52,38	CCCTACTTGAGAACCTTGCAC	87042	I9008	B73
#32	rev	20	61,73	55,00	CCGCTGATGCTATTGACAGG	244983	T9002	B73
#34	fwd	22	59,83	54,55	GTGAGTGAGGTACCACTGCTTG	206676	N9002	B73
#35	fwd	20	61,59	55,00	AACGGTGAGCACTGCTTGAG	207744	O9002	B73
#36	rev	21	62,63	57,14	GCTTAGCCTCTTGTGCTGCTG	208049	O9002	B73
#37	rev	20	61,79	55,00	CCCAATTCGAGGAACTGGTC	208726	P9002	B73
#38	fwd	20	58,67	55,00	GGTGAGGAGGAGAAATCGTC	208579	P9002	B73
#39	rev	20	61,00	60,00	CCTCGTCGAGTCCAACTCTG	209005	P9002	B73
#40	fwd	21	60,82	52,38	GATGTCGCGCAGATAACAGAC	209754	Q9002	B73
#41	rev	21	62,09	52,38	CTTGTTCTCGGAACGAAGGTG	210057	Q9002	B73
#42	rev	22	60,63	50,00	TGGCTGTTTCTCCTAGATGGAC	111057	S9009	Mo17
#43	rev	22	62,31	54,55	GGGCTACTAAGCAGGCAATGTC	208077	O9002	B73
#44	fwd	20	61,95	60,00	GGCTGGAGTGCCAGATTACC	207770	O9002	B73
#45	fwd	18	61,28	61,11	CCGACTGTTCCCATGCAG	180805	G9002	B73
#46	rev	18	59,37	66,67	GTAGGTGCCGAGCCAGAC	180953	G9002	B73
#47	rev	22	61,29	50,00	GTGCCGAATCTAGAGTTGCTTG	181652	H9002	B73
#48	fwd	21	59,59	52,38	TTAGCTCCTGAGCCTGATGAC	181581	H9002	B73
#49	rev	23	58,55	39,13	CAACCATGAAAACGATAGAACAC	181689	H9002	B73
#50	fwd	22	60,87	45,45	GGACAGAAACATGGAAGATTGC	184109	I9002	B73
#51	rev	23	57,34	34,78	ATAAACTGAAATCCATTGAGTCG	184210	I9002	B73
#52	fwd	22	62,32	54,55	CATCCTCTCTCCATGCTTAGGG	187329	J9002	B73
#53	rev	20	62,26	60,00	CTGTCCTTGGCGAGAGATCC	187529	J9002	B73
#54	fwd	20	58,62	55,00	CTCGCAGGACTACAAGATGC	196437	K9002	B73
#55	rev	21	59,72	57,14	TAGTCAGGTAGGGGGCTTCTC	196640	K9002	B73
#56	fwd	20	60,80	55,00	CGAGGCAAACTATGGCAGAC	197400	L9002	B73
#57	rev	20	60,37	55,00	AGATCCAGACTCGGCAGATG	197610	L9002	B73
#58	fwd	22	60,73	50,00	CATAGGCTCAAGGTATGGGTTG	199135	M9002	B73
#59	rev	21	59,68	52,38	GGATGTCATTGCTCTCCAGAC	199291	M9002	B73
#70	fwd	20	60,40	55,00	CTGTGGAAGTTGGGGTTGTC	208967	P9002	B73
#71	fwd	20	60,97	55,00	TGGGGTTGTCAGAGTTGGAC	208977	P9002	B73
#72	rev	20	62,64	55,00	TCAGATGCCGAGAACGTCTG	209789	Q9002	B73
#73	rev	20	61,11	55,00	GTGCACAAGTGGGTGTTGTG	97	AZM_123277	-

Table 5: List of all the tested primer and their features. len: length; tm: melting temperature (°C); Start pos: primer 5' starting position on the contig sequence.

DNA extraction

B73 and Mo17 genomic DNA was extracted using the DNeasy Plant Maxi Kit (Qiagen, Valencia, CA - Cat. No. 68161). DNA quality was checked on agarose gel and concentration was measured by OD_{260}.

The oat-maize addition lines have been produced by Kynast and coworkers from the interspecific cross between oat (*Avena sativa* L.) and maize (*Zea mays* L.) (Kynast et al. 2001). Oat/maize chromosome addition lines genomic DNA (10ng/μl) produced by E. Ananiev were available at the Crop Genetics Group at DuPont Experimental Station (Wilmington, DE).

mRNA purification and reverse transcription

In order to produce biological replicates, the entire procedure from total RNA extraction to cDNA synthesis was independently carried out on 2 different tissue collections for each tissue for each genotype. Further, each extraction was performed starting from tissues blended from 5 individuals, as to minimize individual variations.

Total RNA was extracted using TRIzol Reagent (Invitrogen, Carlsbad, CA - Cat. No. 15596-018) according to the manufacturer protocol. Poly-adenylated RNA was then purified from 500-1000μg of total RNA using the Qiagen Oligotex Midi Kit (Qiagen, Valencia, CA - Cat. No. 70042). 2μg of Poly(A^+)-RNA were then treated with Deoxyribonuclease I (DNase I Amplification Grade, Sigma-Aldrich, Saint Louis, MO - Cat No. AMP-D1) as reported by manufacturer. A PCR control reaction (see below for conditions) on a 2μl aliquot of DNase-treated Poly(A^+)-RNA was performed in order to exclude the possibility of genomic DNA contamination. The primers H2A_fwd (5'agggggtaaggggaagaagg3') and H2A_rev (5'ctacgcatcaccgcatactg3'), designed to the Histone 2A mRNA complete cds sequence [gi:473602] and previously tested on genomic DNA, were used.

The remaining DNAse-treated Poly(A^+)-RNA (about 1.8μg) was then directly retrotranscribed to cDNA using the Reverse Transcription System (Promega, Madison, WI - Cat. No. A3500) using random primers included in the kit. The reactions were setup as indicated by the manufacturer protocol, scaling the total reaction volume up to 40μl (as for 2μg RNA); the 1st strand synthesis step was prolonged to 60 minutes. To verify the cDNA synthesis reaction quality, 6μl of retrotranscription product from 1.2kb Kanamycin Positive Control RNA were analyzed on 1% agarose gel in 1X TBE buffer.

Genomic and RT-PCR

Both genomic and RT-PCR reactions were performed using the HotStarTaq Master Mix Kit (Qiagen, Valencia, CA - Cat. No. 203445) in a final reaction volume of 15μl (7.5μl HotStarTaq Mix 2X; 1μl fwd primer 10μM; 1μl reverse primer 10μM; 2μl template; water to volume). Template consisted of 20ng of genomic DNA for genomic PCR; 2μl of cDNA product diluted 1:5 v/v in sterile ddH$_2$O, as indicated by the manufacturer, were used for RT-PCR. Genomic DNA from oat and from the maize donor line were also tested with the same primer sets respectively as negative and positive controls for the oat-addition lines PCR survey. Thermal cycling were carried out in a GeneAmp® PCR System 9700 thermo-cycler (Applied Biosystems, Foster City, CA, USA), with the following run: a starting step of 10 min. at 95°C, followed by 35 cycles of 45 sec. at 95°C, 45 sec. at 60°C, 1 min. 30 sec. at 72°C; and a final extension of 7 min. at 72°C. Annealing temperature and elongation time have been adjusted on a single-case base, depending on the features of the primers and the dimension of the expected amplification product. PCR results were visualized on 1.5 to 2% agarose gel (ethidium bromide staining) in 1X TBE buffer.

Cloning

Genomic PCR and RT-PCR products were purified using QIAquick PCR Purification Kit (Qiagen, Valencia, CA - Cat. No. 28104) and cloned into pGEM-T Easy Vector (pGEM-T Easy Vector System I (Promega, Madison, WI - Cat. No. A1360) according to the manufacturer protocol. Transformation was performed either by electroporation [1μl ligation + 20μl ElectroMAX DH10B cells (Invitrogen, Carlsbad, CA - Cat. No. 18290-015)] or by heat-shock [5μl ligation + 50μl Chemically Competent *E. Coli* (Invitrogen, Carlsbad, CA - Cat. No. 44-0301)]. Positive transformants were selected by white/blue screening (LB + amp + IPTG + X-Gal). For each PCR product, 12 to 24 positive colonies were picked and inoculated in 96-well format liquid culture (LB-glycerin freezing medium). Positive clones were then confirmed by PCR on 1μl of the liquid culture (same conditions as above), using pUC/M13_for and pUC/M13_rev primers. Results were visualized loading 2μl of PCR product on 1.2% agarose gel (ethidium bromide staining) in 1X TBE buffer.

Sequencing

Depending on DNA concentration, 1 to 4μl of PCR product were cleaned-up using 0.75μl ExoSAP-IT (Amersham Biosciences, Little Chalfont, UK - Cat. No. 78201) in

25.5μl total reaction volume (25 min. 37°C + 15 min. 80°C). The ABI Prism® BigDye™ Terminator v3.1 Ready Reaction Cycle Sequencing Kit (Applied Biosystems, Foster City, CA – Cat. No. 433745x) was used for sequencing [ABI Prism® BigDye™ v3.1, 0.98μl; 5X Sequencing Buffer, 1.6μl; Primer (6.4μM), 1μl; Exo-SAP reaction, 11μl; nuclease-free water to a final volume of 20μl]. Sequencing thermal cycling was carried out in a GeneAmp® PCR System 9700 thermo-cycler (Applied Biosystems, Foster City, CA, USA), with the following run: 25 cycles of 10 sec. at 96°C, 5 sec. at 56°C, 4 min. at 60°C. The 3100 Genetic Analyzer (Applied Biosystems, Foster City, CA) was used for sequencing.

References

Birchler, J.A., Auger, D.L. and Riddle, N.C. (2003). "In search of the molecular basis of Heterosis." The Plant Cell **15**: 2236-2239.

Blake, C. (1983). "Introns - present from the beginning?" Nature **306**: 535-537.

Brosius, J. (1991). "Retroposons--seeds of evolution." Science **251**: 753.

Brunner, S., Fengler, K., Morgante, M., Tingey, S. and Rafalski, A. (submitted). Evolution of DNA sequence non-homologies among maize inbreds.

Crow, J.F. (1948). "Alternative hypotheses of hybrid vigor." Genetics **33**: 477-487.

Davenport, C.B. (1908). "Degeneration, albinism and inbreeding." Science **28**: 454-455.

Doolittle, W.F. (1978). "Genes-in-pieces: were they ever together?" Nature **272**: 581-582.

Dunn, C.A., Medstrand, P. and Mager, D.L. (2003). "An endogenous retroviral long terminal repeat is the dominant promoter for human 1,3-galactosyltransferase 5 in the colon." PNAS **100**: 12841-12846.

Duvick, D.N. (2001). "Biotechnology in the 1930s: the development of hybrid maize." Nat Rev Genet **2**(1): 69-74.

East, E.M. (1908). "Inbreeding in corn." Rep Connecticut Agric Exp Stn for 1907**:** 419-428.

Fu, H. and Dooner, H.K. (2002). "Intraspecific violation of genetic colinearity and its implications in maize." PNAS **99**(14): 9573-9578.

Fu, H., Park, W., Yan, X., Zheng, Z., Shen, B. and Dooner, H.K. (2001). "The highly ricombinogenic *bz* locus lies in an unusually gene-rich region of the maize genome." PNAS **98**(15): 8903-8908.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A. and Smith, O.S. (2004). "Allelic Variation of Gene Expression in Maize Hybrids." Plant Cell **16**(7): 1707-1716.

Kynast, R.G., Riera-Lizarazu, O., Vales, M.I., Okagaki, R.J., Maquieira, S.B., Chen, G., Ananiev, E.V., Odland, W.E., Russell, C.D., Stec, A.O., Livingston, S.M., Zaia, H.A., Rines, H.W. and Phillips, R.L. (2001). "A Complete Set of Maize Individual Chromosome Additions to the Oat Genome." Plant Physiol. **125**(3): 1216-1227.

Lamkey, K.R. and Edwards, J.W. (1999). The quantitative genetics of heterosis. In: Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA**:** 31-48.

Langham, R.J., Walsh, J., Dunn, M., Ko, C., Goff, S.A. and Freeling, M. (2004). "Genomic duplication, fractionation and the origin of regulatory novelty." Genetics **166**: 935-945.

Llave, C., Kasschau, K.D., Rector, M.A. and Carrington, J.C. (2002). "Endogenous and Silencing-Associated Small RNAs in Plants." Plant Cell **14**: 1605-1619.

Long, M., Betràn, E., Thornton, K. and Wang, W. (2003). "Origin of new genes: glimpse from young and old." Nat Rev Genet **4**: 865-875.

Meyer, R.C., Torjek, O., Becher, M. and Altmann, T. (2004). "Heterosis of Biomass Production in Arabidopsis. Establishment during Early Development." Plant Physiol. **134**(4): 1813-1823.

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W. and Fraser, P. (2004). "Active genes dynamically colocalize to shared sites of ongoing transcription." Nat Genet **36**: 1065-1071.

Prince, V.E. and Pickett, F.B. (2002). "Splitting pairs: the diverging fates of duplicated genes." Nat Rev Genet **3**: 827-837.

Ramakrishna, W., Emberton, J., Ogden, M., SanMiguel, P. and Bennetzen, J.L. (2002). "Structural Analysis of the Maize Rp1 Complex Reveals Numerous Sites and Unexpected Mechanisms of Local Rearrangement." Plant Cell **14**(12): 3213-3223.

Rozen, S. and Skaletsky, H.J. (2000). Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S, Misener S (eds) Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ**:** 365-386.

Schnell, F.W., Cockerham, C.C. and Zeng, Z.B. (1996). "Design III with marker loci." Genetics **143**: 1437-1456.

Schramke, V. and Allshire, R. (2003). "Hairpin RNAs and retrotransposon LTRs effect RNAi and chromatin-based gene silencing." Science **301**: 1069-1074.

Shull, G.H. (1908). "The composition of a field of maize." Rep Am Breeders Assoc **4**: 296-301.

Song, R. and Messing, J. (2003). "Gene expression of a gene family in maize based on non-collinear haplotypes." Proc Natl Acad Sci USA **100**: 9055-9060.

Stuber, C.W., Lincoln, S.E., Wolff, D.W., Helentjaris, T. and Lander, E.S. (1992). "Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers." <u>Genetics</u> **132**: 823–839.

Wang, W., Haijing, Y. and Long, M. (2004). "Duplication-degeneration as a mechanism of gene fission and the origin of new genes in *Drosophila* species." <u>Nat Genet</u> **36**: 523-527.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q.F. and Saghai Maroof, M.A. (1997). "Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid." <u>PNAS</u> **94**: 9226–9231.

Extensive *cis*-acting regulatory variation and expression overdominance in maize: a molecular basis for heterosis

Federica Cattonaro[1], Stefano Amadeo[1], Giorgio Pea[2], M. Enrico Pè[2], and Michele Morgante[1]

[1]Università degli Studi di Udine, Dipartimento di Scienze Agrarie ed Ambientali - via delle Scienze 208, 33100 Udine, Italy; [2]Università degli Studi di Milano, Dipartimento di Scienze Biomolecolari e Biotecnologie – via Celoria 26, 20133 Milano, Italy.

Abstract

Maize (*Zea mays*) is known to harbor considerable levels of genetic variation. Nucleotide diversity within single copy regions, including genes, is more than double the interspecific polymorphism rate in mouse. A lack of collinearity in intergenic regions due to the presence of different LTR-retrotransposons has recently been reported and seems to be a widespread phenomenon even between elite US inbreds, making it a unique example of diversity within species. Maize is also characterized by pronounced heterosis (hybrid vigor) that is displayed by the F_1 progenies of crosses between two inbred lines, which has been the basis for the success of maize hybrid. The genetic and molecular basis of heterosis, despite a long history of successful exploitation in many plant species, often also through interspecific crosses, is still unknown. Here we analyzed allele-specific differences in gene expression arising from *cis*-regulatory variation in a random set of genes in two reciprocal maize hybrids. We show that 75% of the genes show allelic differences in expression of at least ±1.5 fold, that differences are tissue specific and that expression overdominance, i.e. a different allele being most highly expressed in different tissues, can also be observed. Besides representing an important source of phenotypic and quantitative variation, regulatory variation may also provide a possible molecular explanation of the heterosis phenomenon in maize. In fact, the patterns of *cis*-regulatory variation we observed in a sample of maize genes are compatible with both dominance and overdominance, which have been in turn proposed as the major genetic cause of the heterotic phenomenon.

Introduction

The development of novel methods for the quantification of allele-specific expression has allowed unveiling the relatively frequent occurrence of differential expression that is not due to imprinting phenomena but most likely to *cis*-acting regulatory variation. Several examples have been reported for humans and mouse, showing that such differences are heritable and context specific (Cowles et al. 2002; Yan et al. 2002; Bray et al. 2003; Pastinen et al. 2004). Allelic diversity has been recently proposed to be an important genetic component for phenotypic variation, especially in plants (Doebley and Lukens 1998; Buckler and Thornsberry 2002; Birchler et al. 2003). Comparative data are also revealing that nucleotide sequences variation widely exists not only between, but also within species. Noticeably, maize genome has revealed an extremely high level of DNA sequence polymorphism, which has been estimated an order of magnitude higher than that observed in human (Sunayaev et al. 2000; Bhattramakki et al. 2002; Buckler and Thornsberry 2002; Ching et al. 2002). The differences range from single nucleotide polymorphisms (SNPs) to large regions of several kilobases (Fu and Dooner 2002). The maize genome is composed largely by repetitive sequences, most of which (70% of the genome) are represented by LTR-retrotransposons (SanMiguel et al. 1996; SanMiguel et al. 1998). Recent reports have suggested that maize inbreds differ largely in the composition of intergenic regions because of the presence of different types of retroelements (Fu and Dooner 2002; Song and Messing 2003; Brunner et al. submitted). Unlike in humans, where most of the genome expansion dates to more than 50 Myrs ago (International Human Genome Sequencing Consortium, 2001), retrotransposon amplification in maize is a fairly recent phenomenon, mostly restricted to the last 3 Myrs (SanMiguel et al. 1998). It is possible that retrotransposon insertions have occurred in the lineages that make up current maize after their divergence from a common ancestor, providing for much greater divergence in intergenic regions than in genic ones. However, divergence in genic regions is still considerable, with nucleotide diversity that is 10 times higher than in humans both for coding and non-coding regions, whenever these are shared and alignable (Sunayaev et al. 2000; Bhattramakki et al. 2002). Lack of collinearity in intergenic regions and high nucleotide diversity are still observed when only the US breeding lines, which are the foundation of modern maize hybrids, are considered (Ching et al. 2002; Fu and Dooner 2002). The interest for maize stems from its specific features in terms of genome and sequence diversity organization as well as from it being a model for

the study and understanding of the heterotic phenomenon. Recently a relationship between heterosis in maize and differential allelic expression resulting from different regulatory regions (*cis* elements) has been proposed (Guo et al. 2004).

We set out to estimate the frequency and magnitude of allele-specific differences in expression levels in a random set of maize genes. In particular, in order to highlight possible tissue-specific regulation, allelic expression ratio for 12 genes were measured in different tissues collected from the F_1 hybrids derived from the reciprocal crosses between the maize inbred lines B73 and H99 (B73xH99 and H99xB73). The implications of tissue-specificity of allelic expression regulation as it might contribute to heterosis are discussed.

Results and Discussion

The analysis of allele-specific expression levels was carried out by comparing two alleles in identical conditions, to control for *trans* and environmental effects, through the analysis of individuals that are in heterozygous condition. We developed a method to detect differences in expression levels due to allelic effects in maize hybrid lines, based upon recent papers about techniques for the precise measurement of allelic levels of gene expression in human and mouse (Cowles et al. 2002; Yan et al. 2002). The transcripts from each of the two alleles were distinguished using single nucleotide polymorphisms (SNPs) in the transcripts themselves, which most likely are not the regulatory mutations but are just used as proxies for them (Figure 1).
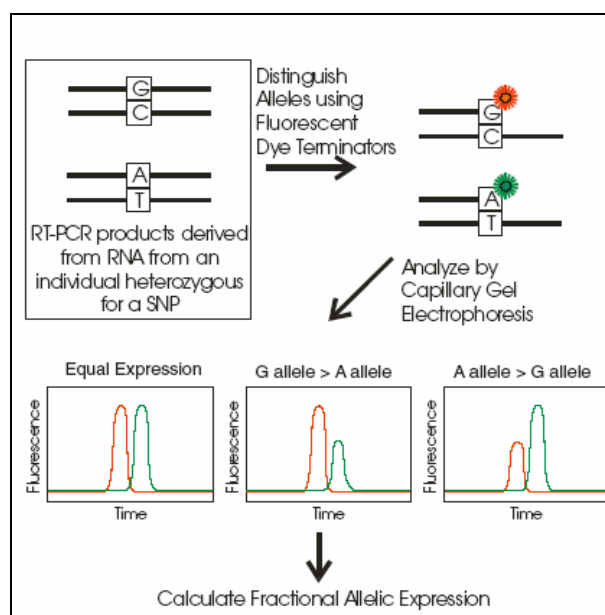
Figure 1: Scheme of the method for the detection of relative differences in allelic expression levels (adapted from Yan et al., 2002). See Material and Methods for more details.

We chose to analyze reciprocal F_1 hybrids (B73XH99 and H99XB73) that were obtained from crossing two standard maize inbreds, belonging to two different heterotic groups, Iowa Stiff Stalk Synthetic (B73) and Lancaster Sure Crop (H99). Crosses between inbred lines from different heterotic groups are the basis for modern maize hybrids. The hybrids we analyzed present pronounced heterosis in terms of grain yield and plant height (Table 1).

Genotype	Yield
B73	5.02
H99	1.86
B73xH99	10.48
H99xB73	12.43
F1 mean	11.45
Mid-parent heterosis	8.96
Best-parent heterosis	6.43

Table 1: Heterosis of the tested hybrids in terms of grain yield (tons/ha; 15.5% humidity; means from 3 plots x 2 replicas x 3 environments). Mid-parent heterosis: (F_1 mean)-(Parents mean); Best-parent heterosis: (F_1 mean) - (Best parent).

We initially identified a set of 57 genes among those that show sequence polymorphisms between inbreds B73 and H99 and that are supposedly expressed in young seedlings. PCR amplifications were performed making sure that introns were not included in the amplified products. Whenever a satisfactory PCR product was obtained, corresponding to single-locus amplification, the presence of the SNPs was confirmed by resequencing of the parental inbreds DNA. Single base extension (SBE) primers were then designed for at least one SNP within each gene. Proportionality between SNP relative signal intensity and ratio of the two alleles was tested using inbred DNA mixtures of known ratios to show linearity of the assay in response to varying concentrations of the two alleles and thus demonstrate the quantitative nature of the assay (data not shown). This also provided a titration curve to normalize measured intensities from RNA samples to those of DNA samples of known relative concentrations. Among the all tested, linearity of the response was obtained for 12 genes, which formed the basis for all further analyses (Table 2). All genes were initially assayed in seedling RNA, and then tested in all remaining tissues whenever an RT-PCR product was obtained.

Acronym	Description	Accession B73
AP1	Similar to (AP001383) hypothetical protein	AI948312
L5	60S Ribosomal protein L5	AI855292
EXPC	Expressed protein C	BM380157
GLIC	Glyceraldehyde-3-P dehydrogenase sub.A	AI973443
AP2	Similar to (AP002063) hypothetical protein	BM080212
ABA	ABA-and ripening-inducible-like protein	BM073855
CHLPR	chlorophyll a/b-binding protein precursor	BM499167
ACP	stearoyl-ACP desaturase	AF498436
PPDK	pyruvate,orthophosphate dikinase (PPDK1)	BQ619338
AT1G	Similar to At1g15980/T24D18_8	BM074154
AMI	protamine	BM073686
PSI	PSI type III chlorophyll a/b-binding protein	BQ539202

Table 2: List of maize genes used for allele-specific expression measurements.

For each tissue, with the exception of immature ears, both reciprocal hybrids were analyzed to be able to distinguish between allele-specific expression due to regulatory variation (the same allele is more highly expressed in the reciprocal hybrids) and due to genomic imprinting (a different allele is more highly expressed in reciprocal hybrids, depending on which allele came through the female or male gamete). A significance threshold of ±1.5 fold for the expression ratio (0.40 or 0.60 if expressed as a proportion on the total transcripts) was chosen based on previous results and on our own analysis of repeated assays.

In seedlings, where all genes appeared to be expressed at detectable levels, 6 out of 12 genes show significant differences in expression levels between the two alleles. Similar results (differences of similar magnitude and favoring the same allele) are observed in the two reciprocal hybrids, ruling out genomic imprinting as a possible cause for the observed differences. Ratios range from 1.5 for ACP to 4.9 fold for EXPC and AP2. In 4 out of the 12 genes we analyzed additional SNPs within the transcript (1 to 3 depending on the gene) that always gave concordant results with the initial analysis and expression ratios of the same magnitude and direction (Table 3). In order to confirm the quantitative nature of the SNP genotyping assay, we also cloned and sequenced RT-PCR products for the *GLIC* gene from seedling RNA. The allelic proportion was 0.67 for B73xH99 and was 0.69 for H99xB73 in comparison to means of 0.65 and 0.61 respectively from the SBE assay.

Gene	Expression proportion (B73 allele)	B73xH99	H99xB73	SNP B73/H99
AP1	A/A+G	0.51 (2;3)	0.59	A/G
L5	G/G+C	0.52 (2;3)	0.56 (2;3)	G/C
EXPC	G/G+C	0.17 (2;3)	0.18 (2;3)	G/C
GLIC	G/G+C	0.65 (2;3)	0.61 (2;3)	G/C
AP2	G/G+A	0.28 (3;4)	0.17 (3;4)	G/A
ABA (SNP1)	C/C+A	0.42	0.54	C/A
ABA (SNP2)	C/C+G	0.57 (2;2)	0.45 (2;2)	C/G
CHLPR (SNP1*)	C/C+T	0.34	0.32 (1;2)	C/T
CHLPR (SNP2*)	G/G+A	0.32	0.23	G/A
ACP	A/A+G	0.59	0.61 (2;2)	A/G
PPDK (SNP1)	G/G+C	0.55	n.d.	G/C
PPDK (SNP2)	C/C+T	0.52	0.53	C/T
AT1G	T/C+T	0.47	0.57	T/C
AMI (SNP1*)	T/C+T	0.8	n.d.	T/C
AMI (SNP2*)	A/G+A	0.73	n.d.	A/G
AMI (SNP3)	A/G+A	0.8	n.d.	A/G
PSI	C/C+T	0.52 (3;4)	0.50 (3;4)	C/T

* SBE primer constructed on the complementary strand; n.d.: not detected

Table 3: Allelic variation in gene expression measured at 12 genes in seedlings mRNAs from the two reciprocal hybrids B73/H99 and H99/B73. The proportion for the B73 allele is always shown. An expression proportion of 0.5 means an identical number of transcripts originating from the two alleles; an expression proportion of 0.4 or 0.6 corresponds to a ±1.5 fold difference in transcript levels from one or the other allele. When multiple measurements were available, numbers in brackets indicate respectively the number of RT-PCR and SBE reactions from which an average value was calculated. For 4 loci the data were confirmed testing additional SNPs (*ABA, CHLPR, PPDK* and *AMI*).

Gene	Expression proportion (B73 allele)	Leaves B73xH99	Leaves H99xB73	Ears B73xH99	Root B73xH99	Root H99xB73	Kernel B73xH99	Kernel H99xB73
AP1	A/A+G	0.37 (2;4)	0.39 (2;4)	0.48 (2;4)	0.49 (1;2)	0.51	0.80 (2;4)	0.19 (2;4)
L5	G/G+C	n.d.	n.d.	n.d.	0.60 (1;2)	0.61 (1;2)	n.d.	n.d.
EXPC	G/G+C	0.3	0.35	0.37 (1;2)	0.50 (1;2)	0.46 (1;2)	0.7	0.16
GLIC	G/G+C	0.75 (1;4)	0.78 (1;4)	0.81	1.00	1.00 (1;2)	1.00	0.00 (1;2)
AP2	G/G+A	0.42 (2;6)	0.37 (2;6)	0.76 (2;6)	0.34 (1;2)	0.36 (1;2)	0.57 (3;5)	0.10 (3;5)
ABA	C/C+G	n.d.	n.d.	0.38 (2;3)	0.47 (1;2)	0.53 (1;2)	1.00 (2;6)	0.00 (1;2)
CHLPR	C/C+T	0.41 (2;4)	0.35 (2;5)	0.40 (1;2)	n.d.	n.d.	0.47	n.d.
ACP	A/A+G	n.d.	n.d.	n.d.	0.65 (1;2)	0.63 (1;2)	n.d.	n.d.

n.d.: not detected

Table 4: Allelic expression measured at 8 genes in 4 different tissues mRNA from the two reciprocal hybrids. The proportion for the B73 allele is always shown. When multiple measurements were available, numbers in brackets indicate respectively the number of RT-PCR and SBE reactions from which an average value was calculated.

We then assayed the same set of genes on a variety of RNAs derived from other tissues or organs, corresponding also to different life-cycle phases. These included roots from young seedlings, leaves from plantlets, immature ears from adult plants and developing kernels (Table 4). Kernels are to be considered separately since they are the only sample derived from a triploid tissue. Due to the double fertilization phenomenon, the endosperm tissue in Angiosperm plants has a triploid genome, two copies of which are of maternal origin and one of paternal one. The expectation for the allelic balance in case of equal expression levels would therefore be of 0.67 and 0.33 for the maternally and paternally derived alleles respectively. The kernels we analyzed, however, although mostly made of triploid endosperm, also contain diploid tissues such as embryo and various seed coat tissues. Unlike for diploid tissues, where the observed allelic expression levels could be compared to an expected 50:50 ratio, a precise expectation cannot be built for our kernel samples. A qualitative and relative expectation though clearly exists, with the maternally derived allele being favored in each of the two reciprocal hybrids, thus providing for a convenient test of the allele expression assay reliability. Four genes (*AP1*, *EXPC*, *GLIC* and *AP2*) could be assayed in all tissues while four others could only be assayed in a sub-sample of them due to lack of detectable expression. Three additional genes with a significant allelic expression ratio difference are detected, namely *AP1* (in leaves), *ABA* (in immature ears) and *L5* (in roots) bringing the total to 9 out of 12 (75%). Despite the lower number of tested genes, this percentage largely exceeds the 10-20% and the 6% values previously observed in human and mouse, respectively (Cowles et al. 2002; Yan et al. 2002; Bray et al. 2003; Pastinen et al. 2004). These results are in agreement with the large lack of collinearity recently observed between different maize inbreds (Fu and Dooner 2002; Song and Messing 2003; Brunner et al. submitted).
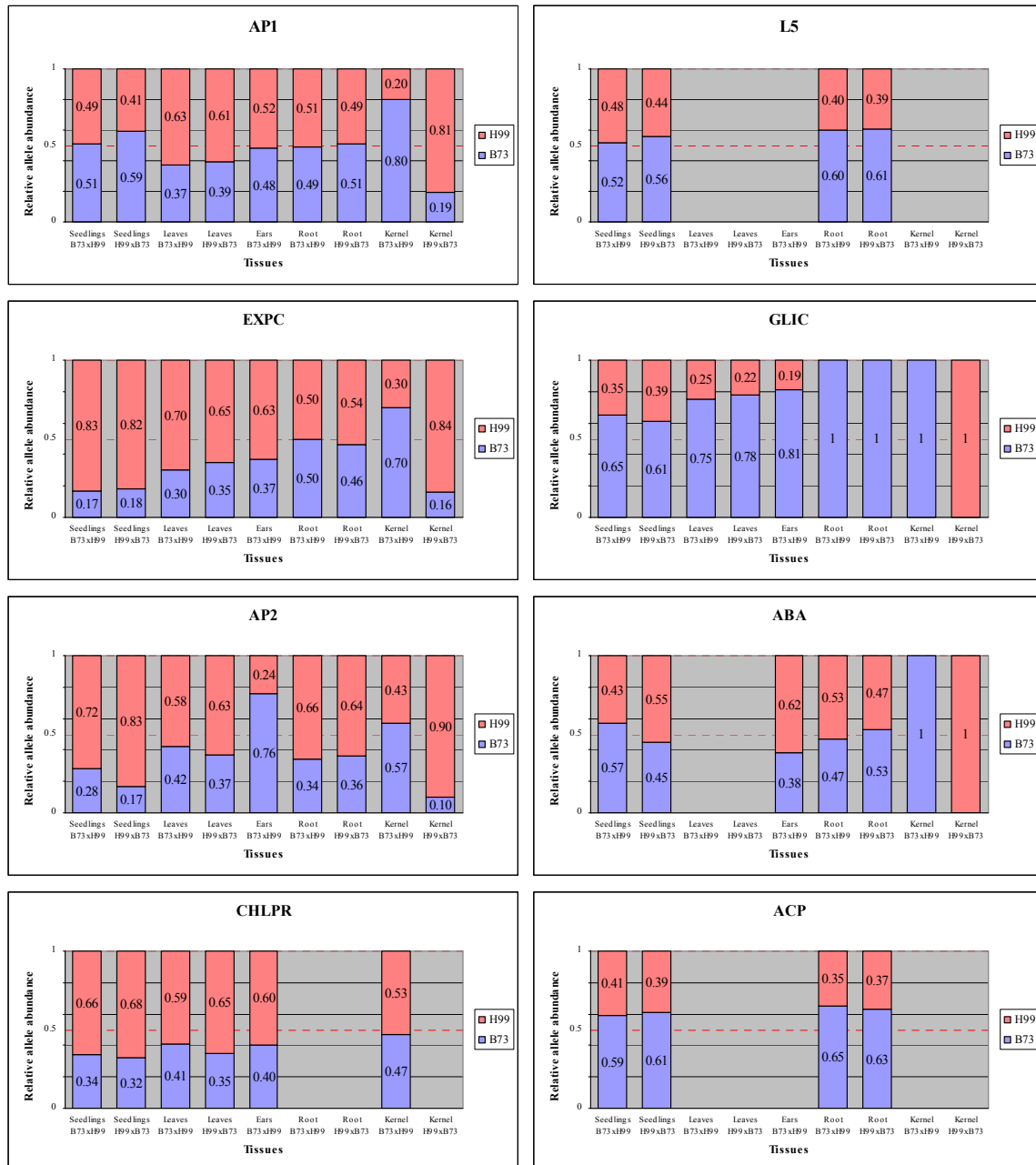
Figure 2: Histograms representing the relative abundance of the alleles in the transcripts from 8 genes tested on different tissues in the B73xH99 and H99xB73 reciprocal hybrids. The dashed line corresponds to the 50:50 ratio expected for equal number of transcripts from the two alleles; an expression proportion of 0.40 or 0.60 corresponds to ±1.5-fold absolute difference between the alleles. With the exception of kernels tissues (see text) variation from the 50:50 ratio indicates allelic preferential expression. In particular, the maintenance of allelic levels consistently different from the 50:50 ratio between the reciprocal hybrids indicate *cis*-regulation effects; while an inversion in relative allelic abundance between the reciprocal hybrids indicates possible imprinting effects. Variation in allele abundance levels among different tissues indicates tissue-specific allelic expression regulation. Absence of bar indicate non detected transcripts (n.d.).

Allelic expression ratios also vary considerably among tissues for the same gene (Figure 2). *EXPC* shows a highly distorted expression ratio in seedlings, favoring the H99 allele, but a 50:50 ratio in roots. The allelic ratios in kernels are in agreement with the triploid genome composition of the endosperm, with the maternal allele always being favored. We detected the expression of a single allele for two genes (*GLIC* and *ABA*) in kernels. The expressed allele is different in the two reciprocal hybrids and always corresponds to the maternally derived one. This provides support for genomic imprinting in the kernel for these genes, a phenomenon that has already been described for the maize endosperm as well as for the Arabidopsis young embryo (Berger 2004; Lauria et al. 2004). *GLIC* gene always shows a higher expression level for the B73 allele; the same allele is exclusively expressed in roots in both the reciprocal hybrids, thus in this case ruling out imprinting as a possible cause. The expression of a single allele is of course in all cases to be considered with caution due to the possibility that a low expression level of the alternative allele may have gone undetected because of the assay sensitivity limits. Finally in three genes, *AP2, AP1* and *ABA*, we observe that a different allele is more highly expressed depending on the tissue we consider. In *AP2* and *ABA* the significant change is detected in the immature ear sample, which is the only sample we analyzed containing reproductive organs/tissues.

Our results confirm recent observations by Guo and coworkers reporting the alleles of 11 out of the 15 analyzed genes (73%) as differentially expressed in two different maize hybrids (Guo et al. 2004), with imprinting effects generally ruled out as negligible. In the same study, when compared over different stressful environments, the less performing hybrid frequently showed mono-allelic expression while the best hybrid tended to express both alleles. Furthermore, relative allelic expression in the two hybrids was found responding differently to abiotic stresses in the two tested tissues (seedlings and ears), suggesting a possible correlation between hybrid performance and allele regulation. Our analysis, spacing on a broader array of different tissues, allowed us to better estimate the extent of tissue-specificity of allelic regulation. In addition to what previously observed, our results show that differential allelic regulation is also occurring within the same hybrid, even when grown in standard conditions. This suggests that a unique expression regulation, resulting from the extremely ductile expression patterns which can exclusively be established in a heterozygous background, might be an ordinary feature of hybrids. Given the fact that most organs we analyzed are made of multiple tissues and cell types, it is likely that the actual departure from equal expression ratios is even more dramatic that the

ones we found in the tested samples. The presence of different regulatory variants with unequal functional properties, inherited from each parent, as well as their combinations, may allow the hybrids responding in different ways according to the cellular or natural environments, thus consistently contributing to hybrid vigor. In other words, allelic regulation, extensively affecting the transcriptome of hybrids, might result in a wide-ranging buffering effect on the regulation of thousands of genes, which in turn would warrant the hybrids a broader range of adaptability. The molecular features of this phenomenon might also account for the observed correlation between parental genetic distance and hybrid performance in maize (Melchinger 1999), as well as for the fact that heterosis in maize culminates at an optimum of parental genetic distance before declining again (Moll et al. 1965). Finally, even if our observations suggest that allelic regulation might primarily act in an overdominant manner, it is also compatible with dominant, pseudo-overdominant and epistatic mechanisms of action, all of which have been invoked as being involved in the determination of heterosis. Further analysis on a larger number of genes, as well as a detailed surveying of the allelic response to different biotic and abiotic stresses, might shed light on both the general and the gene-specific correlation between allelic regulation and heterosis in maize hybrids.

Materials and Methods

Plant material

Plant material was collected from both reciprocal F$_1$ hybrids obtained from the cross between inbred lines B73 and H99 (B73xH99 and H99xB73) and from the two parental inbreds. Seedlings and roots were obtained from plants grown in growth chamber, whereas immature ears and developing kernels were collected from plants grown in open field. In particular, for the former, seeds were sown in 4 cm diameter well plateaus on inert substrate (Agriperlite, BPB Italia SpA, Italy) and incubated in a growth chamber with a 14 hours light/10 hours dark photoperiod at 26° C. Plant material was collected from each individual according to its developmental stage, when the third adult leaf apex became visible. The term seedling refers to any vegetative tissues above the first adult leaf, which was excluded. From the same individuals the entire root apparatus was collected and substrate debris was washed off by quick immersion in distilled water. At least ten individual plantlets were pooled for the root and seedling RNA extraction. Leaves were obtained from plants grown in a greenhouse under natural light. Seeds were sown in soil

and leaf blades collected 20 days after the first leaf emerged from the plumule. A single leaf from a plant was used for the leaf RNA extractions. Immature ears were collected from ears in which well developed silks were present in the basal half of the ear. Silks were discarded, together with the top half of each ear. Developing kernels were collected 15 days after pollination. In particular, the two reciprocal F_1 kernels were produced by cross-pollination between the two parental inbred lines, either one used as pollen donor or as female. First internode of appropriate plants was horizontally severed, and plant placed in water and within minutes carried to the bench where kernels were separated with a cutter. In order to minimize environmental effects, for each tissue plant material was collected approximately at the same time of the day, frozen in liquid nitrogen and stored at –80°C until used. To average individual variation, for each genotype material from different plants was bulked prior RNA extraction.

RNA Extraction and purification

Frozen tissues were grinded in liquid nitrogen and total RNA was purified with TRIZOL (Invitrogen, Carlbad, CA) following manufacture protocols, except for an additional 5 minutes centrifugation in TRIZOL reagent and an additional chloroform extraction. Total RNA was resuspended in DEPC-treated mQ water and stored at –80°C. Quality of RNA was tested by electrophoresis on 1.5% agarose gel and quantified by absorbance at 260nm. Total RNAs extracted with Trizol Kit were treated with Deoxyribonuclease I (Amplification Grade, Sigma) prior to reverse transcription. RNAs from young leaves (10 days old) were extracted using the SV Total RNA Isolation Kit (Promega) according to the manufacturer's protocol.

Allele expression assay

Publicly available SNPs between maize inbred lines B73 and H99 were identified for 35 loci (http://www.ncbi.nlm.nih.gov; http://maize.math.iastate.edu/isumaize/). 22 additional ESTs sequences from TIGR Maize Database (http://www.tigr.org/tdb/tgi/maize/) were chosen from B73 cDNA libraries '5955' (from shoot) and '#8MT' (from seedlings and silk). The 22 PCR products from TIGR EST contigs were resequenced in B73 and H99 to identify SNP tags for allele specific gene expression analyses. PCR primers that flanked the marker polymorphism were designed by using the Primer3 program (www-genome.wi.mit.edu/cgi-bin/primer/primer_3www.cgi). SBE primers were designed with a minimum length of 18nt.

Out of 57 loci, 12 were chosen for allele-specific gene expression assays. The remaining were discarded on the basis of a) putative low expression level in seedlings or unknown codified product; b) absent, poor or aspecific PCR amplification; c) absence of the putative SNPs; d) failure of the SBE reaction; e) lack of linearity of signals in known mixtures of genomic DNA of B73 and H99.

PCR amplifications on genomic DNA were carried out in a total reaction volume of 25 µl containing 1X GenAmp PCR Buffer II (Applied Biosystems), 1% DMSO, 2mM MgCl$_2$, 200 µM dNTPs, 1µM of primer mix (forward plus reverse), 1.25 U of Taq Gold (Applied Biosystems), 20 ng of DNA. Thermocycling conditions consisted of an initial denaturation step of 95°C for 10 min., followed by 38 cycles of 94°C for 1 min., 55°C for 1 min., 72°C for 1 min., with a final extension step at 72°C for 7 min. RT-PCR reactions were performed using the Access RT-PCR system (Promega; 60°C annealing temperature, 47 cycles for cDNA amplification). Genomic and RT-PCR products were verified by agarose gel electrophoresis. Amplified samples from RT-PCR and PCR reactions were incubated with Exo-SAP IT (Amersham Biosciences) according to the manufacturer's instructions prior to primer extension reaction. Primer extension was carried out with the MegaBACE SNuPe Genotyping Kit (Amersham Biosciences). Reactions were performed in a total volume of 10 µl, containing 5 µl of treated PCR or RT-PCR product diluted 1:10, 4 µl of SNuPe premix, 1 µl of 2 µM SNP-specific primer. Primer extension thermocycling conditions consisted of 20 cycles of 96°C for 10 sec., 60°C for 10 sec. and 50°C for 10 sec. Following primer extension, reactions products were purified with Multiscreen 96-well HV Filtration plates (Millipore) loaded with Sephadex G-50 Superfine (Sigma). Aliquots of 5 µl of SNuPe reaction product were combined with 5 µl of MegaBACE loading solution (Amersham Biosciences) and 0,025 µl of MegaBACE SNuPe Multiple Injection Marker (Amersham Pharmacia) and loaded on a MegaBACE 500 capillary sequencer (Amersham Biosciences). Peak heights representing allele-specific extended primers were determined by the software MegaBACE Genetic Profiler v. 2.0 (Amersham Biosciences). The ratios between peak heights were expressed as B73/(B73+H99) peak heights. Mixes of genomic DNA of B73:H99 were prepared in proportion 1:1, 3:1, 1:3 and SBE reactions on these templates were run alongside with the cDNA and no RT control samples. The genomic mixes allowed the construction of a titration curve by linear regression from which the ratio for the RNA samples was extrapolated. The obtained ratios were normalized on the basis of the peak height ratio measurements obtained from SBE on hybrid genomic DNA, representing a perfect 50:50 ratio of the two alleles.

Determination of the relative level of mRNA accumulation in hybrids by direct sequencing

RT-PCR and PCR products obtained with the Glyceraldehyde-3-phosphate dehydrogenase (GLIC) locus-specific primers were cloned from the two hybrids B73xH99 (BH) and H99xB73 (HB) using the p-GEM-T Easy System II Kit (Promega). The reconstitution of the pool of transcripts was made by sequencing of random clones for each reciprocal cross for PCR and RT-PCR products (88 sequences from genomic BH, 92 from cDNA BH, 87 sequences from genomic H99xB73 and 78 cDNA sequences from H99xB73).

Sequencing

Sequencing reactions of PCR and cloned products were carried out using the Big-Dye Terminator Cycle Sequencing kit v. 2.0 (Applied Biosystems) and run on an Applied Biosystems 3730 DNA Analyzer (Applied Biosystems).

References

Berger, F. (2004). "Imprinting - a green revolution." Science **303**: 483-485.

Bhattramakki, D., Dolan, M., Hanafey, M., Wineland, R., Vaske, D., Register, J.C., Tingey, S. and Rafalski, A. (2002). "Insertion-deletion polymorphisms in 3' regions of maize genes occur frequently and can be used as highly informative genetic markers." Plant Mol Biol **48**: 539-547.

Birchler, J.A., Auger, D.L. and Riddle, N.C. (2003). "In search of the molecular basis of Heterosis." The Plant Cell **15**: 2236-2239.

Bray, N.J., Buckland, P.R., Owen, M.J. and O'Donovan, M.C. (2003). "*Cis*-acting variation in the expression of a high proportion of genes in human brain." Hum Genet **113**: 149-153.

Brunner, S., Fengler, K., Morgante, M., Tingey, S. and Rafalski, A. (submitted). Evolution of DNA sequence non-homologies among maize inbreds.

Buckler, E.S. and Thornsberry, J.M. (2002). "Plant molecular diversity and applications to genomics." Curr Opin Plant Biol **5**: 107-111.

Ching, A., Caldwell, K.S., Jung, M., Dolan, M., Smith, O.S., Tingey, S., Morgante, M. and Rafalski, A. (2002). "SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines." BMC Genet **3**: 19.

Cowles, C.R., Hirschborn, J.N., Altshuler, D. and Lander, E.S. (2002). "Detection of regulatory variation in mouse genes." Nat Genet **32**: 432-437.

Doebley, J. and Lukens, L. (1998). "Transcriptional regulators and the evolution of plant form." Plant Cell **10**: 1075-1082.

Fu, H. and Dooner, H.K. (2002). "Intraspecific violation of genetic colinearity and its implications in maize." PNAS **99**(14): 9573-9578.

Guo, M., Rupe, M.A., Zinselmeier, C., Habben, J., Bowen, B.A. and Smith, O.S. (2004). "Allelic Variation of Gene Expression in Maize Hybrids." Plant Cell **16**(7): 1707-1716.

Lauria, M., Rupe, M.A., Guo, M., Kranz, E., Pirona, R., Viotti, A. and Lund, G. (2004). "Extensive maternal DNA hypomethylation in the endosperm of *Zea mays*." The Plant Cell **16**: 510-522.

Melchinger, A.E. (1999). Genetic diversity and heterosis. In: Coors JG and Pandey S (eds) "The Genetics and Exploitation of Heterosis in Crops." ASA, CSSA and SSA: Madison, WI, USA**:** pp 99-118.

Moll, R.H., Lonnquist, J.H., Fortuna, J.V. and Johnson, C.E. (1965). "The relationship of heterosis and genetic divergence in maize." Genetics **52**: 139-144.

Pastinen, T., Sladek, R., Gurd, S., Sammak, A.a., Ge, B., Lepage, P., Lavergne, K., Villeneuve, A., Gaudin, T., Brandstrom, H., Beck, A., Verner, A., Kingsley, J., Harmsen, E., Labuda, D., Morgan, K., Vohl, M.-C., Naumova, A.K., Sinnett, D. and Hudson, T.J. (2004). "A survey of genetic and epigenetic variation affecting human gene expression." Physiol. Genomics **16**(2): 184-193.

SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y. and Bennetzen, J.L. (1998). "The paleontology of intergene retrotransposons of maize." Nat Genet **20**: 43-45.

SanMiguel, P., Tikhonov, A., Jin, Y.K., Motchoulskaia, N., Zakharov, D., Melake-Berhan, A., Springer, P.S., Edwards, K.J., Lee, M., Avramova, Z. and Bennetzen, J.L. (1996). "Nested retrotransposons in the intergenic regions of the maize genome." Science **274**: 765-768.

Song, R. and Messing, J. (2003). "Gene expression of a gene family in maize based on non-collinear haplotypes." Proc Natl Acad Sci USA **100**: 9055-9060.

Sunayaev, S.R., Lathe, W.C., Ramensky, V.E. and Bork, P. (2000). "SNP frequencies in human genes: An excess of rare alleles and differing modes of selection." Trends Genet **16**: 335-337.

Yan, H., Yuan, W., Velculescu, V.E., Vogelstein, B. and Kinzler, K.W. (2002). "Allelic variation in human gene expression." Science **297**: 1143.

Acknowledgments

Well… for the next acknowledgments, I guess I must skip back to English. In the first place, I want to vividly thank Antoni for the fabulous opportunity he gave me to live and enjoy a great experience in his lab, as well as for the absolutely enjoyable time we spent in great discussions. It's really been a pleasure for me. Thanks for everything!

Also many thanks with no distinction to all the components of its group, for having been patient, kind and absolutely helpful in everything. Many many thanks to Scott for welcoming me in Crop Genetics Group at DuPont, and of course thanks to all the people at DuPont Experimental Station who helped me at any time. I would especially like to mention Diana here, whom I've too often bothered with my own businesses…

A big thank to all the great guys I've met at Cosmopolitan Club too, for having made my staying here a little bit easier, and certainly funnier! And to all my room-mates in Greenville apartment, Anne-Celine, Katrin and Mai, for the enjoyable companionship.

Finally, I must propose a standing ovation for Stephan, to whom I literally owe my survival in America! He shouldn't have changed the tires of his car so soon if it hadn't been for me… We worked well together, but first of all he's been a great fellow and a good friend. And always will be. Thanks for patience, for discussions, for dinners and brunches, for travels, for laughs, for coffee breaks and bla bla bla™ (this is all me, isn't it?). As well as just for sitting in front of good beers. Thanks! …actually I also owe you an apology: your addiction to caffeine is my entire fault!

Infine, più importante di qualunque altro, un enorme GRAZIE va ai miei genitori, senza i sacrifici, le rinunce e il sostegno dei quali queste pagine non sarebbero mai state scritte.