Development and Application of Functional Markers

Thomas Lübberstedt, Yongsheng Chen, Everton A. Brenner, Imad Zein

Department of Agronomy, Iowa State University,
1204 Agronomy Hall, Ames, IA, 50011-1010
THOMASL@IASTATE.EDU

Introduction

The value of anonymous genetic markers such as random DNA markers (SSRs, AFLPs, RFLPs etc.) depends for various applications on the known linkage phase between marker and target locus alleles (Lübberstedt et al. 1998). Thus, (quantitative) trait locus mapping is necessary for each cross *de novo*, as different subsets of QTL are polymorphic in individual populations, and linkage phases between marker and QTL alleles can disagree even in closely related genotypes. In contrast, "functional markers" (FMs) (Andersen and Lübberstedt 2003) are a good "translator" from genomic technologies into improved crop varieties. FMs are derived from polymorphic sites within genes, quantitative trait nucleotides (QTN) or quantitative trait insertion – deletion mutations (QTINDEL), causally affecting phenotypic trait variation. Once genetic effects have been assigned to functional sequence motifs, FMs can be used for fixation of beneficial alleles. FM development requires (1) functionally characterized genes, (2) allele sequences from such genes, (3) identification of polymorphic, functional motifs affecting plant phenotype within these genes, (4) validation of associations between DNA polymorphisms and trait variation, and (5) conversion into technical assays using, e.g., any of the single nucleotide polymorphism (SNP) or INDEL detection technologies (Syvänen 2001).

Development of Functional Markers: Identification of QTN and QTINDELs

Genomics-Driven Gene Identification

Projects addressing systematic sequencing of the maize genome contributed large amounts of publicly available sequences during the past few years. More than 2 Mio. expressed sequence tag (EST) sequences have been released to the public domain (http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html). The maize genome sequencing project is nearing completion (http://dev.maizesequence.org/index.html). These millions of sequence reads provide an excellent basis for gene identification by (i) sequence homology, (ii) synteny to rice or other grasses, and (iii) forward genetic approaches (e.g., map-based gene isolation). Especially map-based approaches will benefit from the systematic sequencing of BAC contigs planned for the next years. Major resources for community-based large scale map-based characterization of agronomic traits are the NAM population (Yu et al. 2008) and the PVP-derived materials (Johnson 2007).

Further high-throughput tools have been established to identify candidate genes for traits of interest, such as microarrays (http://www.maizegdb.org/microarray.php, http://www.maizearray.org/). Traditionally, maize transposons play an important role as tool in maize gene discovery. Comprehensive resources mainly based on the Mutator transposon have been and are being established (http://www.mutransposon.org/project/) both for forward and reverse genetic screening of traits and genes, respectively. In conclusion, the prospects for identification of genes affecting agronomic characters can be expected to increase substantially within the next decade.

Allele Sequencing
Whereas resequencing has long been very costly and laborious, recent progress in sequencing technology allows to generate billions of bp at affordable costs (e.g., Margulis et al. 2005). Using this technology, major portions of the genomes of the 25 NAM founder lines in addition to B73 will be sequenced (http://www.maizegenetics.net/).

QTN and QTINDEL Identification
Assignment of an "agronomic function" to short sequence motifs can be achieved by candidate gene based association studies (Risch 2000). This approach is limited by linkage disequilibrium (LD), i.e. haplotype structures in the gene(s) of interest. However, for several genes a generally low LD was detected in maize (Flint-Garcia et al. 2003), including examples in elite materials (Zein et al. 2007). Thus, candidate gene-based association studies are promising in maize. In heterogeneous genotype collections associations identified for specific sites might be confounded with effects from other genome regions especially in case of population stratification (Pritchard et al. 2000), which needs to be taken into account for interpretation of results from association studies.

In a pioneering study, Thornsberry et al. (2001) demonstrated the feasibility of association studies in maize to identify sequence polymorphisms within genes affecting characters of agronomic significance. While taking population structure into account, nine SNP or INDEL polymorphisms were shown to significantly affect flowering time in a set of 92 diverse maize lines. In part, these results were confirmed in a collection of European elite inbred lines (Andersen et al. 2005, Veyrieras et al. 2007). The major reason for non-significance of some of the nine polymorphisms identified by Buckler very likely was the much narrower genetic material investigated by Andersen et al. (2005) as compared to Thornsberry et al. (2001).

In vivo validation of QTN and QTINDELs
Isogenic lines allow to test the impact of particular chromosome segments on a trait of interest. However, introgressions typically span several cM, increasing the chance of confounding effects from adjacent genes. Moreover, even if by use of large populations and tightly linked markers introgressions would be confined to the target gene (allele), it would still not be possible to separate and test effects of individual sequence features within genes. *Transformation* of various alleles of a particular gene of interest into the same background would eliminate the risk of confounding effects caused by linkage drag. However, transgenic events are vulnerable to position effects, which substantially affect the expression of genes depending of the (random) introgression site in the genome.

A public *TILLING* resource has been established for maize (http://genome.purdue.edu/ maizetilling/index.htm). By use of TILLING, various isogenic mutants within a gene of interest can be obtained. However, given an average size of maize genes of maize genes of 10 kb, and that particular SNP alleles are required for validating the outcomes of association studies, it is very unlikely to find mutant genotypes exactly matching the findings from association studies.

Homologous recombination (HR) is the ideal method for *in vivo* validation of QTN and QTINDELs identified by association studies, based on allele replacement. HR is routine in yeast, mice, and Physcomitrella, but occurs at very low frequencies in higher plants (Terada et al. 2007). If it becomes available as routine approach, it will be a major breakthrough for QTN and QTINDEL validation.

Finally, differences between maize “macro-alleles” (Brunner et al. 2005), with frequent insertions or deletions even of complete genes might impact the expression of the target gene. Salvi et al. (2007) demonstrated for *Vgt1*, that regulatory elements might be about 100 kb away from the expressed part of the gene.

Development of Functional Markers: Technical aspects for conversion into assays

Over the past years, a large number of methods based on different techniques and chemistries of allele discrimination, reaction format, and detection platforms have been developed (Syvänen 2001). SNP detection relies on four different mechanisms: (1) allele specific hybridization, (2) allele specific nucleotide incorporation (primer extension), (3) allele specific oligonucleotide ligation, and (4) allele specific invasive cleavage. Determining the most appropriate SNP detection method depends upon technical features of the method in relation to application scenarios, the throughput-level, and costs (both in term of capital investment and costs per assay).

Most SNP detection methods can be automated and offer the possibility of automatic genotype scoring. Their capacity for multiplexing and the labour costs per data point determine efficiency and throughput (Holloway et al. 1999; Törjék et al. 2003; Pati et al. 2004; Giancola et al. 2006). Four popular methods representing different chemistry, detection platform, and multiplex level were compared in Table 2. Real-time (RT) PCR using a variety of detection systems, e.g., TaqMan has become an established method due to its simple, fast and high throughput format (Holloway et al. 1999). TaqMan PCR technology relies on allele-specific hybridization of a probe carrying a fluorescent reporter molecule at the 5’ end, and a quencher molecule at the 3’ end. Upon cleavage by the 5' exonuclease activity of Taq polymerase during PCR, the reporter dye will fluoresce as it is no longer quenched and the intensity of the emitted light is measured. Modified probes such as LNA (Locked Nucleic Acids), a modified nucleic acid analogue, showed better hybridization properties than standard TaqMan probes (Kennedy et al. 2006). TaqMan is a simple assay, since all reagents are added at once into a reaction well in 96- or 384 well formats. PCR and data calling occur simultaneously (real time mode) with average project time of 2 hours per run including 15-20 % hands-on time. Even through the TaqMan assay is performed as monoplex or duplex, one person is able to generate up to 2000 data points per day in a monoplex, due to its simplicity. When the assay is run in end point mode (only suitable for some applications) and duplex, 3000 data points are possible. TaqMan has shown to be a suitable technique for GMO tests (Iida et al. 2005), MAS (Kennedy et al. 2006), and MAB (Helguera et al. 2003).

SNaPshot® (Applied Biosystems, Foster City, USA) is based on mini-sequencing, i.e., single-base extension using fluorescent labelled dideoxy nucleotides (ddNTPs). The size of the product is the size of the initial probe/primer plus one fluorescent base. For multiplexing, primers are designed in different lengths of 23 to 60 bases with 4 to 5 nucleotide differences. The reactions may be carried out in 5- to 10-plexes using capillary electrophoresis for data detection in 96 well format (Törjék et al. 2003). The SNaPshot assay includes multiple steps (Pati et al. 2004). Total project time from PCR to data evaluation is on average 40 minutes per 10-plex, with hands-on time of approximately 15% (corresponding to 6 minutes) of the total project time. Using multi-capillary electrophoresis (96 capillaries), one person can generate >10.000 data points per day.

MassARRAY® (Sequenom, San Diego, USA) is based on single-base extension and follows the same overall principles as the SNaPshot assay. However, MassARRAY uses standard ddNTPs and MALDI-TOF MS (Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry) for marker detection. 40-plex reactions in 384 well format allow a single person to generate >100 000 data points per day (Bagge and Lübberstedt 2008). MassARRAY is flexible and suitable to generate both small and large marker numbers per sample (Jones et al. 2007).

Another high-throughput method is the GoldenGate assay (Illumina Inc, San Diago, USA) offering a 96, 384, 768 or 1536plex custom selected SNP assay allowing 96 genotypes to be assayed in parallel (Fan et al. 2003). Allele-specific primers are hybridised directly to genomic DNA immobilised on a solid support. In case of a perfect match the primer is extended and the extension product is ligated to a probe hybridised downstream the SNP position. The ligated product is amplified by PCR using primers that are complementary to a sequence in the 3'-end of the ligation probes and 5'-ends of the allele-specific primers, respectively. After PCR, the amplified products are captured on beads carrying complementary target sequences for the SNP-specific Tag of the ligation probe. In this assay format up to 300 000 data points can be generated by one person per day.

Examples in Maize

Pioneering association studies in maize were performed by the group of Dr. Buckler (e.g., Thornsberry et al. 2001). Meanwhile, association analyses and QTN / QTINDEL discovery were conducted for various traits including starch biosynthesis (Wilson et al. 2004), flowering time (Thornsberry et al. 2001, Salvi et al. 2007), carotinoid biosynthesis (Palaisa et al. 2003, Harjes et al. 2008), and morphological characters (Botiri et al. 2006).

Several candidate genes in relation to cell-wall digestibility and forage quality are meanwhile available (http://www.polebio.scsv.ups-tlse.fr/MAIZEWALL/index.html). Varying levels of LD have previously been observed between genes of the phenylpropanoid pathway, decaying within few hundred bps for *CCoAOMT2* and *COMT* (Guillet-Claude et al. 2004a, b, Zein et al. 2007) while spanning more than 3.5 kb at the *PAL* locus (Andersen et al. 2007). In heterogeneous genotype collections associations identified for specific sites might be confounded with effects from other genome regions especially in case of population stratification (Pritchard et al. 2000), which needs to be taken into account for interpretation of results from association studies.

First reports on association studies for genes involved in cell wall biosynthesis confirm that these pathways are promising targets for identification of polymorphic sites associated with forage quality, and thus FM development (Barriere et al. 2003). Zein et al. (2007) investigated the sequence variation at the *Bm3* locus in a collection of 42 European maize inbred lines, contrasting with respect to stover DNDF and relevant for hybrid maize breeding in Central Europe. For association with forage quality, stover digestibility was determined in six environments between 2001 and 2003 in Germany (heritability >0.9). One INDEL polymorphism within the intron revealed significant association with stover digestibility (Lübberstedt et al. 2005). In a study of Guillet-Claude et al. (2004a), polymorphisms both in the AldOMT (= *Bm3*) and the CCoAOMT2 but not CCoAOMT1 coding genes showed significant association with maize digestibility. The *PAL* gene was investigated in a set of 32 European elite inbred lines (Andersen et al. 2007). A one-bp deletion in the second exon of *PAL*, introducing a premature stop codon, was associated with high IVDOM. Moreover,

polymorphisms in the maize peroxidase gene ZmPox3 were also significantly associated with maize digestibility (Guillet-Claude et al. 2004b). In conclusion, availability of qualified candidate genes can be effectively converted into informative molecular markers by means of association studies.

Challenges in Functional Marker Development

Linkage Disequilibrium

Even in populations with substantial intragenic LD decay, adjacent polymorphic sites might still be in complete LD, obscuring the unequivocal identification of causal QTN or QTINDEL polymorphisms. This will, in consequence, lead to an overestimation of trait-associated polymorphisms. Another LD-related issue is the identification or development of optimal QTN or QTINDEL haplotypes, if several polymorphisms within the target gene affect the trait of interest. If not available in the characterized population, development of optimal QTN or QTINDEL allele combinations based on intragenic recombination events might be difficult to achieve, even by use of large populations and intragenic markers. Alternatively exotic germplasm might provide a source for novel intragenic combinations of QTN and QTINDEL alleles.

Consistency of Functional Markers

Comparable to QTL studies, a major concern after detection of QTN or QTINDELs in association studies is transferability of information gained in one study to other situations. Apart from the question of false positives, transferability of true QTN or QTINDELs might be affected by the composition of populations in different studies, both with regard to allele frequencies at the target locus, and structure of the respective populations. Dwarf 8 is the only example in plants so far, where the same locus has been studied independently in different experimental populations of inbred lines (Thornsberry et al. 2001, Andersen et al. 2005, Charcosset et al. 2007). Although the QTN and QTINDELs identified by Thornsberry et al. (2001) were not significant in the study of Andersen et al. (2005), they were confounded with the structure of the population used in the latter study. Other factors with potential impact on the detection of QTN or QTINDELs are epistatic and dominance (so far, association studies in maize were conducted at line *per se* level), as well as environment and genotype by environment effects. In conclusion, the genetic effects of QTN or QTINDELs are background-, population-, and environment dependent. Thus, presence of a beneficial QTN or QTINDEL allele merely reflects a certain potential of trait expression, analogous to the risk concept in human genetic diseases, depending on the genetic effect and penetrance of the respective allele.

Pleiotropy

For application of functional markers, it will be essential to test for negative pleiotropic side-effects. This will in addition lead to a better understanding of the nature of trait correlations, or "pleiotropic" effects described for major genes. Various studies found close genetic correlations between plant height and flowering time. Interestingly, the flowering time associated polymorphisms in Dwarf 8, a gene initially identified by its mutant allele leading to dwarfing, had no effects on plant height (Thornsberry et al. 2001, Andersen et al. 2005). Similarly, mutant alleles of brown midrib genes in maize were found to affect other agronomic characters including plant height and biomass yield (Pedersen et al. 2005). However, none of the polymorphisms within the *Bm3* gene affecting forage quality, affected any of these agronomic traits (Chen et al., unpublished data). In conclusion, composition of

optimal haplotypes for genes shown to affect one or more traits of interest needs to take multiple traits into consideration.

Application of functional markers

Given further progress in sequencing technology providing large amounts of sequence data at low costs, sequencing of additional maize inbreds beyond B73 will soon become reality. Projects like the NAM community approach (Yu et al. 2008) will lead to accumulation of further characterized genes and QTN / QTINDELS of agronomic relevance. Thus, the amount of functionally characterized polymorphisms in maize as prerequisite for functional marker development will substantially increase over the next decade. Functional markers might be useful for various steps along the process of cultivar development. These include (1) identification of novel or better alleles (QTN / QTINDEL haplotypes) for characterized genes in exotic germplasm collections, (2) identification of complementary parents prior to generation of genetic variation for development of new inbreds, (3) description of the "genetic potential" of new inbreds, and (4) variety registration and description. It remains to be seen, how functional markers contribute to marker-assisted (recurrent) selection, in particular as compared to genomic selection procedures based on low cost markers without requirements on their functional characterization (Bernardo 2007). Calus et al. (2008) showed that haplotype versus random marker-based genomic selection is more efficient to predict breeding values. It appears likely, that marker-multiplexes employed in genomic selection procedures based on previously characterized QTN or QTINDELs are at least superior to random markers in populations with low linkage disequilibrium.

References

Andersen JR, Lübberstedt T (2003) Functional markers in plant. Trends Plant Sci 8:554-560

Andersen JR, Schrag T, Zein I, Melchinger AE, Lübberstedt T (2005) Functional marker validation of polymorphisms in the maize dwarf8 gene affecting flowering time in European elite materials. Theor. Appl. Genet. 111: 206-217

Andersen JR, Zein I, Krützfeldt B, Eder J, Ouzunova M, Wenzel G, Lübberstedt T (2007) Polymorphisms at the Phenylalanine Ammonia-Lyase locus are associated with forage quality in European maize (*Zea mays* L.). Theor Appl Genet 114:307-319

Bagge M, Lübberstedt T (2008) Functional markers in wheat – technical and economic aspects. Mol Breed 22:319-328

Barriere Y, Guillet C, Goffner D, Pichon M (2003) Genetic variation and breeding strategies for improved cell wall digestibility in annual forage crops. A review. Anim. Res. 52: 193-228

Bernardo R, Yu J (2007) Prospects for genomewide selection for quantitative traits in maize. Crop Sci 47:1082-1090

Bortiri E, Chuck G, Vollbrecht E, Rocheford T, Martienssen R, Hake S (2006) ramosa2 encodes a LATERAL ORGAN BOUNDARY Domain Protein That Determines the fate of stem cells in branch meristems of maize. Plant Cell

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17:343–360

Calus MPL, Meuwissen THE, de Roos APW, Veerkamp RF (2008) Accuracy of genomic selection using different methods to define haplotypes. Genetics 178:553-561

Fan J B et al. (2003) Highly parallel SNP genotyping. Cold Spring Harb Symp Quant Biol 68:69-78

Flint-Garcia M, Thornsberry JM, Buckler ES (2003) Structure of linkage disequilibrium in plants. Annu. Rev. Plant Biol. 54:357–74

Giancola S et al (2006) Utilization of the three high-throughput SNP genotyping methods, the GOOD assay, Amplifluor and TaqMan, in diploid and polyploidy plants. Theor Appl Genet 112:1115-1124

Guillet-Claude C, Birolleau-Touchard C, Manicacci D, Fourmann M, Barraud S, Carret V, Martinant JP, Barrière Y (2004a) Genetic diversity associated with variation in silage corn digestibility for three O-methyltransferase genes involved in lignin biosynthesis. Theor Appl Genet 110:126-135

Guillet-Claude C, Birolleua-Touchard C, Manicacci D, Rogowsky PM, Rigau J, Murigneux A, Martinant JP, Barriere Y (2004b) Nucleotide diversity of the ZmPox3 maize peroxidase gene: relationships between a MITE insertion in exon 2 and variation in forage maize digestibility. BMC Genetics 5:19

Harjes CE et al (2008) Natural genetic variation in Lycopene Epsilon Cyclase tapped for maize biofortification. Science 319, 330-334

Helguera M et al (2003) PCR assays for the *Lr37-Yr17-Sr38* cluster of resistance genes and their use to develop isogenic red spring wheat lines. Crop Sci 43:1839-1847

Holloway JW et al (1999) Comparison of three methods for single nucleotide polymorphism typing for DNA bank studies: Sequence-specific oligonucleotide probe hybridization, TaqMan liquid phase hybridization, and microplate array diagonal gel electrophoresis (MADGE). Human Mutation 14:340-347

Iida M et al (2005) Development of Taxon-specific sequences of common wheat for the detection of genetically modified wheat. J Agric Food Chem 53:6294-6300

Jones ES et al (2007) A comparison of simple sequence repeat and single nucleotide polymorphism marker technologies for the genotypic analysis of maize (*Zea mays* L). Theor Appl Genet 115:361-371

Kennedy BG, Waters DLE, Henry RJ (2006) Screening for the rice blast resistance gene *Pi-ta* using LNA displacement probes and real-time PCR. Mol Breeding 18:185-193

Lübberstedt T, Melchinger AE, Fähr S, Klein D, Dally A, Westhoff P (1998) QTL mapping in testcrosses of flint lines of maize: III. comparison across populations for forage traits. Crop Sci 38: 1278-1289

Lübberstedt T, Zein I, Andersen JR, Wenzel G, Krützfeldt B, Eder J, Ouzunova M, Shi C (2005) Development and application of functional markers in maize. Euphytica 146:101-108

Margulies M et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380

Palaisa KA, Morgante M, Williams M, Rafalski A (2003) Contrasting effects of selection on sequence diversity and linkage disequilibrium at two phytoene synthase loci. Plant Cell 15:1795–1806

Pati N et al (2004) A comparison between SNaPshot, pyrosequencing, and biplex invader SNP genotyping methods: accuracy, cost and throughput. J Biochem Biophys Methods 60:1-12

Pedersen JF, Vogel KP, Funnell DL (2005) Impact of Reduced Lignin on Plant Fitness. Crop Sci. 45:812–819

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000) Association mapping in structured populations. Am.J.Hum.Genet 67: 170-181

Risch NJ (2000) Searching for genetic determinants in the new millennium. Nature 405: 847-856

Salvi S et al. (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. PNAS 104: 11376–11381

Syvänen AC (2001) Acessing genetic variation: genotyping SNPs. Nature Rev. Genetics 2: 930-942

Terada R, Johzuka-Hisatomi Y, Saitoh M, Asao H, Iida S (2007) Gene Targeting by Homologous Recombination as a Biotechnological Tool for Rice Functional Genomics. Plant Physiology 144: 846–856

Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001) Dwarf8 polymorphisms associate with variation in flowering time. Nat.Genet. 28: 286-289

Törjék O et al (2003) Establisment of a high-efficient SNP-based framework marker set for *Arabidopsis*. Plant Journal 36:122-140

Veyrieras JP et al. (2007) Bridging Genomics and Genetic Diversity: Linkage Disequilibrium Structure and Association Mapping in Maize and Other Cereals Crop Sci. 47(S3) S60–S71

Wilson LM, Whitt SR, Ibanez AM, Rocheford TR, Goodman MM, Buckler ES (2004) Dissection of maize kernel composition and starch production by candidate gene association. Plant Cell 16:2719–2733

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic Design and Statistical Power of Nested Association Mapping in Maize. Genetics 178: 539–551

Zein I, Wenzel G, Andersen JR, Lübberstedt T (2007) Low level of linkage disequilibrium at the COMT (Caffeic acid O-methyltransferase) locus in European maize (*Zea mays* L.). Gen Res & Crop Evol. 54: 139-148