Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (*Zea mays* L.) germplasm

Elizabeth Jones · Wen-Chy Chu · Mulu Ayele · Julie Ho · Ed Bruggeman · Ken Yourstone · Antoni Rafalski · Oscar S. Smith · Michael D. McMullen · Chethana Bezawada · Jana Warren · Jean Babayev · Sutirtha Basu · Stephen Smith

Received: 20 November 2008 / Accepted: 1 April 2009 / Published online: 22 April 2009

Abstract The development of single nucleotide polymorphism (SNP) markers in maize offers the opportunity to utilize DNA markers in many new areas of population genetics, gene discovery, plant breeding and germplasm identification. However, the steps from sequencing and SNP discovery to SNP marker design and validation are lengthy and expensive. Access to a set of validated SNP markers is a significant advantage to maize researchers who wish to apply SNPs in scientific inquiry. We mined 1,088 loci sequenced across 60 public inbreds that have been used in maize breeding in North America and Europe. We then selected 640 SNPs using generalized marker design criteria that enable utilization with several SNP chemistries. While SNPs were found on average every 43 bases in 1,088 maize gene sequences, SNPs that were amenable to marker design were found on average every 623 bases; representing only 7% of the total SNPs discovered. We also describe the development of a 768 marker multiplex assay for use on the Illumina® BeadArray™ platform. SNP markers were mapped on the IBM2 intermated B73 × Mo17 high resolution genetic map using either the IBM2 segregating population, or segregation in multiple parent-progeny triplets. A high degree of colinearity was found with the genetic nested association map. For each SNP presented we give information on map

Electronic supplementary material The online version of this article (doi:10.1007/s11032-009-9281-z) contains supplementary material, which is available to authorized users.

E. Jones (✉) · M. Ayele · E. Bruggeman ·
K. Yourstone · O. S. Smith · J. Warren ·
J. Babayev · S. Basu · S. Smith
Pioneer Hi-Bred International, Inc. (DuPont Agriculture
and Nutrition), 7300 NW 62nd Avenue,
P.O. Box 1004, Johnston, IA 51031-1004, USA
e-mail: liz.jones@pioneer.com

W.-C. Chu · C. Bezawada
Pioneer Hi-Bred International, Inc. (DuPont Agriculture
and Nutrition), 810 Sugar Grove Ave-Highway 44,
Dallas Center, IA 50063-2005, USA

J. Ho
Pioneer Hi-Bred International, Inc. (DuPont Agriculture
and Nutrition), 1039 S Milton-Shopiere,
P.O. BOX 668, Janesville, WI 53547-0668, USA

A. Rafalski
DuPont Agriculture and Nutrition, Route 141,
Henry Clay Bldg #353, Wilmington,
DE 19880-0353, USA

M. D. McMullen
Plant Genetics Research Unit, USDA-Agricultural
Research Service, 302 Curtis Hall, Columbia,
MO, USA

M. D. McMullen
Division of Plant Sciences, University of Missouri,
Columbia, MO 65211-7020, USA

location, polymorphism rates in different heterotic groups and performance on the Illumina® platform.

Keywords Maize · Single nucleotide polymorphisms (SNP) · Illumina · Expected heterozygosity · Genetic mapping

Introduction

Single nucleotide polymorphism (SNP) markers offer the promise of higher map resolution, higher throughput, lower cost and a lower error rate compared to simple sequence repeat (SSR) markers, thereby enabling new applications in molecular breeding. (Gupta et al. 2001; Bhattramakki et al. 2002; Rafalski 2002; Batley et al. 2003; Vroh Bi et al. 2006; Jones et al. 2007). Whereas standardization of SSR data across laboratories can be challenging due to variation in allele sizing, with SNP markers, the nucleotide itself is usually reported thereby enabling comparisons across laboratories, technologies and chemistries. This feature will facilitate standardization of germplasm identification on a world-wide basis.

SNP discovery and marker development are well underway in maize (*Zea mays* L.). The most significant effort has been the NSF-funded project 'Molecular and Functional Diversity of the Maize Genome' (Zhao et al. 2006; http://www.panzea.org/). In this project, SNP discovery has been performed on over 3,000 genes, with genetic mapping data on over 1,100 SNP markers being collected on 'Nested Association Maps' (NAM) maps using diverse maize inbred lines (Zhao et al. 2006; Yu et al. 2008). Maize has a relatively high frequency of SNPs (Tenaillon et al. 2001; Ching et al. 2002; Vroh Bi et al. 2006) offering an abundance of sites for potential interrogation. However, this feature can also hinder marker design due to polymorphisms in the DNA sequence around the targeted SNP. Further obstacles can occur during SNP validation due to additional polymorphisms present in germplasm that was not included in the initial marker design stage. The success rate of marker validation is greatly improved through having access to deep, high quality sequence alignments across a wide array of relevant germplasm, or else through having information on SNPs that have previously been successfully validated on a wide array of germplasm. Here we describe the criteria and process that resulted in the selection of 640 SNPs discovered in sequences from public inbreds that have been used in maize breeding programs in the US and in Europe. The deposited sequences deepen existing available sequence and can aid in further SNP discovery and successful design. We provide information on the development of a SNP marker set with the Illumina® BeadArray™ platform that can be used for whole genome fingerprinting across a range of maize germplasm. Information is also provided on quality, map location, and expected heterozygosity values to assist researchers in developing additional SNP marker sets.

Materials and methods

SNP discovery

High quality, non-redundant expressed sequence tags (ESTs) were submitted to Myriad Genetics, Inc, Salt Lake City for design of nested primer pairs (two forward and two reverse) directed towards the 300–500 bases flanking the last exon and 3′UTR. The 3′UTR region was sequenced because SNPs had previously been found to be more frequent in this region (Ching et al. 2002). The primers were used to amplify the DNA of 24 diverse Pioneer-proprietary inbreds. Those that gave good amplification were selected for further amplification of additional inbreds of interest. Amplification products were sequenced, the sequences were aligned and SNPs and indels were identified using software tools proprietary to Myriad. The sequences, SNPs and indels were then QCd by human inspection of the alignments and reported using Pioneer-proprietary software (Genomix, Hanafey et al. unpublished). For this study, 1,088 sequence loci were selected that had high quality sequence and showed a match by BLASTN (Altschul et al. 1997) at an initial expected value cutoff of e^{-10} and 100 nt overlap to the public unigene sequences (Cone et al. 2002; Gardiner et al. 2004) contained within the Panzea database (http://www.panzea.org). Sequences were examined for 60 public inbreds that have been used in maize breeding in North America and Europe, consisting of 12 Flint, 22 non-stiff stalk, 19 stiff stalk, 3 tropical and 4 miscellaneous inbreds that together encompass diverse origins (Table 1). Based on consensus sequences for the 1,088 loci, SNP and indel frequencies

Table 1 Maize inbreds sequenced and used for marker design, and used for Illumina marker validation

Maize inbred	Material type	Heterotic grouping	Sequenced[a]	Illumina validation[b]
38-11	Public	Stiff stalk	Y	
A165	Public	Non-stiff stalk	Y	
A188	Public	Miscellaneous	Y	
A509	Public	Non-stiff stalk	Y	Y
A556	Public	Non-stiff stalk	Y	
A619	Public	Non-stiff stalk	Y	
A632	Public	Stiff stalk	Y	
B	Public	Non-stiff stalk	Y	
B14	Public	Stiff stalk	Y	
B37	Public	Stiff stalk	Y	Y
B42	Public	Non-stiff stalk	Y	
B64	Public	Stiff stalk	Y	
B73	Public	Stiff stalk	Y	Y
B84	Public	Stiff stalk	Y	
B89	Public	Stiff stalk	Y	
B94	Public	Stiff stalk	Y	
C103	Public	Non-stiff stalk	Y	
C106	Public	Flint	Y	
CI66	Public	Miscellaneous	Y	
CM49	Public	Flint	Y	
CM7	Public	Flint	Y	
CML197	Public	Tropical		Y
CML312	Public	Tropical		Y
CO109	Public	Flint	Y	
D02	Public	Flint	Y	
D146	Public	Flint	Y	Y
F2	Public	Flint	Y	
F252	Public	Miscellaneous	Y	
F257	Public	Flint	Y	
F283	Public	Flint	Y	
F7	Public	Flint	Y	Y
GT119	Public	Non-stiff stalk	Y	
H84	Public	Stiff stalk	Y	
H99	Public	Non-stiff stalk	Y	
HATO4	Public	Flint	Y	Y
HY	Public	Stiff stalk	Y	Y
I205	Public	Non-stiff stalk		Y
Indiana H60	Public	Non-stiff stalk	Y	
K187-11217	Public	Stiff stalk	Y	
K55	Public	Miscellaneous	Y	
L1546	Public	Non-stiff stalk	Y	
L317	Public	Non-stiff stalk	Y	
Minn49	Public	Non-stiff stalk	Y	
MO13	Public	Miscellaneous	Y	

Table 1 continued

Maize inbred	Material type	Heterotic grouping	Sequenced[a]	Illumina validation[b]
MO17	Public	Non-stiff stalk	Y	Y
MP305	Public	Tropical	Y	Y
N28	Public	Stiff stalk	Y	Y
OH07	Public	Non-stiff stalk	Y	
OH40B	Public	Non-stiff stalk	Y	
OH43	Public	Non-stiff stalk	Y	Y
OH45	Public	Non-stiff stalk	Y	
OS420	Public	Stiff stalk	Y	Y
OS426	Public	Stiff stalk	Y	Y
PA91	Public	Non-stiff stalk	Y	
PH05F	Pioneer	Stiff stalk		Y
PH07D	Pioneer	Stiff stalk		Y
PH09B	Pioneer	Stiff stalk		Y
PH0AV	Pioneer	Non-stiff stalk		Y
PH0PD	Pioneer	Tropical		Y
PH1073	Pioneer	Non-stiff stalk		Y
PH11DT	Pioneer	Tropical		Y
PH11DV	Pioneer	Tropical		Y
PH12A	Pioneer	Tropical		Y
PH161	Pioneer	Non-stiff stalk		Y
PH165	Pioneer	Non-stiff stalk		Y
PH17J	Pioneer	Flint		Y
PH1B5	Pioneer	Non-stiff stalk		Y
PH1CN	Pioneer	Non-stiff stalk		Y
PH1W2	Pioneer	Stiff stalk		Y
PH24E	Pioneer	Non-stiff stalk		Y
PH26N	Pioneer	Tropical		Y
PH290	Pioneer	Non-stiff stalk		Y
PH2N0	Pioneer	Non-stiff stalk		Y
PH350	Pioneer	Stiff stalk		Y
PH475	Pioneer	Non-stiff stalk		Y
PH508	Pioneer	Non-stiff stalk		Y
PH5FW	Pioneer	Flint		Y
PH605	Pioneer	Stiff stalk		Y
PH630	Pioneer	Non-stiff stalk		Y
PH661	Pioneer	Non-stiff stalk		Y
PH680	Pioneer	Tropical		Y
PH695	Pioneer	Non-stiff stalk		Y
PH707	Pioneer	Stiff stalk		Y
PH84D	Pioneer	Non-stiff stalk		Y
PH8CW	Pioneer	Non-stiff stalk		Y
PH953	Pioneer	Stiff stalk		Y
PH9HP	Pioneer	Tropical		Y
PHB72	Pioneer	Tropical		Y

Table 1 continued

Maize inbred	Material type	Heterotic grouping	Sequenced[a]	Illumina validation[b]
PHB89	Pioneer	Non-stiff stalk		Y
PHBD5	Pioneer	Non-stiff stalk		Y
PHBR2	Pioneer	Stiff stalk		Y
PHG29	Pioneer	Non-stiff stalk		Y
PHG49	Pioneer	Flint		Y
PHG63	Pioneer	Tropical		Y
PHHB4	Pioneer	Stiff stalk		Y
PHHB9	Pioneer	Stiff stalk		Y
PHJ40	Pioneer	Stiff stalk		Y
PHJ89	Pioneer	Non-stiff stalk		Y
PHJ90	Pioneer	Non-stiff stalk		Y
PHJRT	Pioneer	Tropical		Y
PHJRV	Pioneer	Tropical		Y
PHK46	Pioneer	Non-stiff stalk		Y
PHK76	Pioneer	Non-stiff stalk		Y
PHKV1	Pioneer	Stiff stalk		Y
PHM10	Pioneer	Non-stiff stalk		Y
PHM26	Pioneer	Tropical		Y
PHM2K	Pioneer	Tropical		Y
PHM49	Pioneer	Non-stiff stalk		Y
PHMK0	Pioneer	Stiff stalk		Y
PHN37	Pioneer	Non-stiff stalk		Y
PHN46	Pioneer	Non-stiff stalk		Y
PHN49	Pioneer	Tropical		Y
PHNV8	Pioneer	Flint		Y
PHP38	Pioneer	Stiff stalk		Y
PHP51	Pioneer	Tropical		Y
PHPDH	Pioneer	Tropical		Y
PHPDJ	Pioneer	Tropical		Y
PHPDK	Pioneer	Tropical		Y
PHR03	Pioneer	Non-stiff stalk		Y
PHR25	Pioneer	Non-stiff stalk		Y
PHR47	Pioneer	Stiff stalk		Y
PHSKB	Pioneer	Tropical		Y
PHT18	Pioneer	Tropical		Y
PHV78	Pioneer	Non-stiff stalk		Y
PHV87	Pioneer	Flint		Y
PHW52	Pioneer	Stiff stalk		Y
PHWVV	Pioneer	Tropical		Y
PHZ51	Pioneer	Non-stiff stalk		Y
R159	Public	Stiff stalk	Y	
SC213R	Public	Non-stiff stalk	Y	
SD105	Public	Stiff stalk	Y	
SRS303	Public	Non-stiff stalk	Y	Y

Table 1 continued

Maize inbred	Material type	Heterotic grouping	Sequenced[a]	Illumina validation[b]
TR9-1-1-6	Public	Non-stiff stalk	Y	
TX601	Public	Miscellaneous	Y	
V3	Public	Flint	Y	
W153R	Public	Stiff stalk	Y	
WF9	Public	Stiff stalk	Y	

[a] Maize inbreds sequenced and used for SNP discovery

[b] Maize inbreds used for Illumina® BeadArray™ SNP marker validation

were estimated. For the SNP loci only, measures of nucleotide diversity were assessed using θ^d per nucleotide (Watterson 1975) and Π^D per nucleotide (Tajima 1983) in Arlequin (Version 3.1; Excoffier et al. 2005).

SNP loci were further selected for marker development using broad marker design criteria to allow utilization with several SNP chemistries. These criteria were: availability of DNA sequence 100 bp upstream and downstream from the SNP site, 25 bp around the target SNP with no polymorphism and a consensus sequence GC content of 40–60%. This resulted in 640 SNPs in 540 sequences being selected. Individual and consensus sequence data representing the 60 public inbreds are available at http://www.panzea.org and are identified by a PHM ('Pioneer Hi-Bred Marker') prefix. The PHM SNP loci are also described in supplementary Table 1.

SNP validation on the Illumina GoldenGate platform

First round of validation (Illumina_1)

The 640 PHM SNP loci, plus an additional 895 'PZA' SNP loci identified from sequencing a diverse range of maize and teosinte [*Zea mexicana* (Scrad.) Kuntze] inbreds (Wright et al. 2005; available from http://www.panzea.org) were submitted to Illumina, Inc., San Diego, California for assessment of marker designability using proprietary criteria. While the PHM markers were all pre-selected to be amenable to design with several marker chemistries (see previous section), the PZA markers were not pre-selected based on any design criteria before being submitted to Illumina. Markers were assigned a designability score by Illumina based on proprietary criteria; 1 (highly designable), 0.5 (moderately designable) or 0 (low designability). Seven hundred and sixty eight SNPs (listed under column 'Illumina_1' in supplementary Table 1) were selected that were greater than 50 nucleotides apart when in the same sequence (as recommended by Illumina design criteria) and that preferentially had a designability score of 1. Markers were validated using the Illumina GoldenGate genotyping assay with BeadArray™ technology using 91 public and Pioneer-proprietary flint, stiff-stalk, non-stiff stalk and tropical maize inbreds (Table 1). Marker success was assessed in terms of allele call frequency (the frequency of alleles successfully called out of the 480 maize inbreds assayed) and by comparison to sequence data for public genotypes B73 and Mo17.

Second round of validation (Illumina_2)

Markers with >70% data and polymorphic in at least one class of public or proprietary stiff stalk, non-stiff stalk, flint or tropical germplasm were selected for the second Illumina multiplex assay design. Those excluded were replaced by PHM markers with high designability (designability of 1), or PZA markers that had previously been validated in Illumina assays in the laboratory of one of the authors (M. McMullen) and were found to be polymorphic across a range of temperate and tropical lines. The second 768 multiplex assay (listed in supplementary Table 1 under the column 'Illumina_2) was tested with newly extracted DNA from the 91 public and proprietary inbreds (Table 1).

This multiplex assay was also used to profile proprietary and public maize samples in four subsequent projects. Information on the percent allele calls found for each project and an overall average is provided in supplementary Table 1.

Evaluation of expected heterozygosities

For the 640 PHM SNPs, expected heterozygosity values [calculated as: $H = 1 - \Sigma(p_i^2)$, where p_i is the frequency of the *i*th allele] were calculated using sequence data for the 60 public inbreds (Table 1). For the markers included in the Illumina validation assays, expected heterozygosity values were also calculated for the 91 public and Pioneer-proprietary inbreds included in the Illumina validation study (Table 1). Expected heterozygosity values were calculated across all inbreds and also by heterotic group (individual marker values in supplementary Table 1; average values in Table 5).

Number of alleles and expected heteorzygosity values were also calculated for each sequence haplotype based on patterns of SNPs within each sequence. All SNPs called by Myriad Genetics were included in assigning the haplotype allele and not just those SNPs that met the generalized design criteria.

Genetic mapping

Two different methods were used to estimate coordinates for PHM and PZA markers on the IBM2 intermated B73 × Mo17 high resolution genetic map (Lee et al. 2002; http://www.maizegdb.org/). In the first method, PHM and PZA markers were used to genotype 284 lines of the public IBM2 population. A two-point linkage analysis was performed for each PHM and PZA marker and approximately 2,000 public SSR markers that appear on the IBM2 map to estimate IBM2 locations.

A second method was used for PHM and PZA markers that were monomorphic in the IBM2 population, or for markers which failed to give satisfactory segregation data in this population. Recombination data obtained from genotyping 439 Pioneer inbred lines that could be grouped according to parent-progeny relatedness were used to place markers on a proprietary Pioneer map that also contains approximately 2,000 public IBM2 markers. PHM and PZA SNP marker locations on the IBM2 map were inferred by co-location with public markers common to the IBM2 and Pioneer proprietary maps.

SNPs residing in the same PHM or PZA sequence as another SNP mapped using one of the above methods were assigned the same map location. PHM SNPs were also aligned with a GenBank accession sequence using BLASTN (Altschul et al. 1997) at an initial expected value cutoff of e^{-10} and 100 nt overlap. This information was used to align PHM with PZA SNPs through their common GenBank accession sequence. SNPs residing in the same GenBank accession sequence as a SNP mapped using one of the above genetic methods were tentatively assigned the same map location.

Map locations were compared to the NAM (nested association maps) map (NAM_map_20080419.xls) published at the Panzea website (http://www.panzea.org/lit/data_sets.html#NAM_map; Yu et al. 2008).

Results

SNP discovery

From the 1,088 genic sequences (consisting of 398,425 bases) targeted for marker design, a total of 2,140 insertions/deletions (indels) and 9,194 SNPs were found using Myriad Genetics criteria (Table 2). This represents, on average, 1 indel every 186 bases and 1 SNP every 43 bases. Average θ per nucleotide was 0.0033 and Π was 0.0036.

Our primary goal was to select a set of SNP loci that would be amenable to design using a number of

Table 2 SNP discovery in public inbred sequences

	Average per sequence	Minimum per sequence	Maximum per sequence	Total for 1,088 sequences selected
Sequence length (bases)	369	160	940	398,425
Indels (insertions/deletions)	2.0	0	24	2,140 (1 per 186 bp)
SNPs	8.5	0	40	9,194 (1 per 43 bp)
SNPs meeting broad marker design criteria[a]	1.2	1	4	640 (1 per 623 bp)

[a] 100 flanking bases upstream and downstream from the SNP site, 25 bp around the target SNP with no polymorphism and a GC content of 40–60% using consensus sequence

different platforms. For this reason, the initial criteria for SNP selection were fairly conservative. These criteria resulted in 640 usable SNPs from 540 sequences (Table 2). This is on average 1 SNP every 623 bases, or 7% of the total SNPs discovered.

SNP validation on the Illumina platform

Illumina_1

The 640 PHM SNPs and 895 PZA SNPs were submitted to Illumina for assessment of designability according to their proprietary criteria. 1,119 (71.7%) SNPs were assessed as being highly designable (Table 3). These represented 89.5% of the PHM loci and 60.6% of the PZA loci. The 30% higher value for the PHM SNPs reflected the pre-selection for generalized marker designability for these markers prior to submission to Illumina.

A set of 768 SNP loci consisting of 484 PHM loci and 292 PZA loci was selected for validation. The majority of SNP markers produced high quality data; 93.2% markers generated allele scores on >70% inbreds and with alleles that agreed with sequence scores (Table 3). The frequency of allele calls is presented for each marker in supplementary Table 1.

Illumina_2

Following the second round of validation, 88.8% of the markers gave expected allele calls for >70% of the samples (Table 4). This represented a lower success rate than for Illumina_1, despite 90% of the SNPs having been previously validated. The lower success rate was possibly due to the inbred DNA being re-extracted for this experiment and potentially being of a lower quality. In addition, markers were re-synthesized for this experiment, so that the two validation experiments are largely independent and cannot be directly compared. The merit of prior SNP validation could be assessed by comparing the previously validated PHM SNPs with the non-validated PHM SNPs in Illumina_2. For the previously successfully validated markers, 10.5% gave poor quality data in Illumina_2 (Table 4). According to Illumina, a 10% failure rate for previously successfully validated SNPs is expected for the Illumina GoldenGate system due to failure during oligonucleotide synthesis and bead array manufacture. New PHM markers that had not previously been tested in an Illumina experiment had double this failure rate at 20.9% (Table 4).

Expected heterozygosity

The set of 640 PHM SNPs collectively had an average expected heterozygosity value of 0.33 across the 60 public genotypes (Table 5). In order to minimize ascertainment bias, Illumina validation markers were not selected based on high polymorphism levels and so had similar average expected heterozygosity values averaging 0.37 (Table 5). Expected heterozygosity

Table 3 PHM and PZA loci submitted for assessment of marker designability

	PHM	PZA	Total
Total submitted for electronic design	640	895	1,560
Designability 1 (high level of designability)	573 (89.5%)	542 (60.6%)	1,119 (71.7%)
Designability 0.5 (moderate designability)	54 (8.4%)	135 (15.1%)	191 (12.2%)
Designability 0 (low designability)	13 (2.0%)	218 (24.4%)	232 (14.9%)

Table 4 Marker success rate in the first and second rounds of Illumina validation

	Marker success rate (>70% alleles called, allele calls agree with sequence data)		
	PHM	PZA	All markers
Illumina_1	92.2% (439/476)	94.8% (277/292)	93.2% (716/768)
Illumina_2	88.3% (391/443)	89.5% (291/325)	88.8% (682/768)
SNPs tested in Illumina_1	89.5% (358/400)	91.4% (192/210)	90.2% (550/610)
SNPs not tested in Illumina_1	79.1% (34/43)	87.8% (101/115)[a]	85.4% (135/158)

[a] These PZA SNPs had previously been successfully validated on the Illumina platform in the lab of M. McMullen

Table 5 Number of alleles and expected heterozygosity across different sets of inbreds for 540 PHM sequence haplotypes, the 640 PHM SNPs selected as candidates for marker design, and all markers (PHM and PZA) in the final Illumina multiplex assay

Sequence or marker set	Average number of alleles (min–max)	Expected heterozygosity (min–max)	
		Public inbred set[a]	Public and Pioneer inbreds[b]
540 PHM sequences	5.9 (2–15)	0.61 (0.03–0.88)	No data
640 PHM SNPs	2	0.33 (0.03–0.5)	No data
443 PHM SNPs in Illumina_2	2	0.37 (0.04–0.5)	0.34 (0–0.50)
All functional SNPs in Illumina_2 (PHM and PZA)	2	No data	0.35 (0–0.50)

[a] The 60 public inbreds used for sequencing and SNP discovery (Table 1)

[b] The 91 public and pioneer-proprietary inbreds used for SNP validation on the Illumina platform (Table 1)

values are presented for each individual SNP within each heterotic group in supplementary Table 1. As an illustration of the effects of ascertainment bias, expected heterozygosities within pairs of heterotic groups were plotted for each marker and very little correlation was found (e.g. Fig. 1).

The average expected heteorzygosity for sequence haplotypes was twice as large as for SNPs (Table 5).

SNP mapping

The majority of PHM SNPs could be placed on the public IBM2 map by one of the two methods described in the Materials and Methods. Only 31 (4.8%) of the 640 PHM SNPs were not mapped. The Illumina multiplex had a larger number of unmapped loci (23%) mainly due to the incorporation of PZA markers for which we had no information in Pioneer inbred populations and so were unable to map using this method. However, 34 of the unmapped PZA markers in the multiplex have map locations on the preliminary NAM map (see supplementary Table 1).

Map locations for a small proportion of the SNP loci were assigned through being in the same GenBank accession sequence as another SNP mapped using genetic mapping methods. To determine the validity of this approach, PZAs and PHMs assigned to the same GenBank accession sequence and that had been independently mapped with genetic mapping methods were examined. Thirteen GenBank accession sequences were found that were associated with more than 1 SNP mapped using genetic methods. Of these, 8 pairs of SNPs had the same map location, 2 had map locations less than 5 cM apart, but 3 had map locations on different chromosomes.

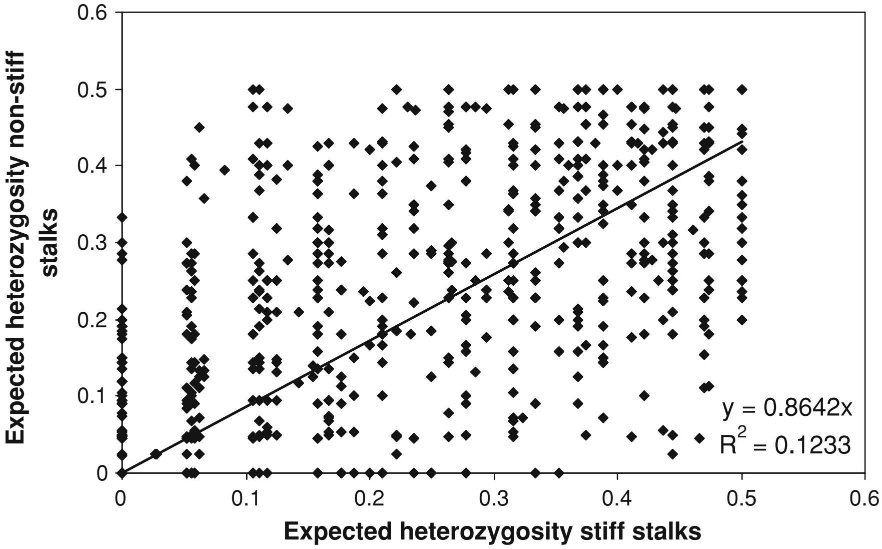

Fig. 1 Correlation between expected heterozygosity values for individual PHM markers among stiff stalks and non-stiff stalk individuals in the public inbred set

A total of 282 markers were common to the IBM2 and NAM maps. Map distances per se could not be directly compared as the IBM2 map represents multiple meioses whereas the NAM map represents a single meiosis. However, when markers common to the two maps were aligned, a high degree of colinearity was found. There were 32 differences in map order, with 30 being within 2.2 cM on the NAM map and one within 8 cM. There was only one marker with a different chromosome assignment; PZA00708.3 mapped to chromosome 9 on the NAM map and chromosome 1 by association with another SNP in the same GenBank accession sequence on the IBM2 population, indicating that marker locations assigned via physical analysis may not agree with those assigned using genetic mapping.

The 591 markers in the Illumina multiplex assay that had IBM2 map locations covered 88% (6,651 cM) of the IBM2 2005 Neighbors map with an average marker spacing of 8.75 cM (range 0–105.9 cM).

Discussion

We present here 640 PHM SNPs that have been selected to be usable with a number of SNP chemistries and that can be utilized in genetic studies and breeding applications. We also provide information on the development of a 768 multiplex assay on the Illumina platform and provide data for each marker concerning polymorphism and success rate. The pre-selection of SNP markers for general designability that we have undertaken will increase the likelihood of these SNPs being successfully validated on any platform. We found that such pre-selection for the PHM loci gave a 30% increase in highly designable markers over the PZA markers that had not been pre-selected using these criteria. We also found that prior information on success rate on the Illumina platform reduced failure rate from 20 to 10%. Many of these SNPs have also previously been tested on the Invader (Third Wave Technologies, Madison, WI, USA) and MassARRAY (Sequenom, San Diego, CA, USA) platforms and their success documented (Jones et al. 2007).

There was very little correlation between expected heterozygosity values determined in the different heterotic groups, substantiating previous observations that ascertainment bias can be introduced when selecting SNPs in a particular germplasm set (Clark et al. 2005; Rafalski and Tingey 2008). Here, we attempted to reduce ascertainment bias through developing SNP markers using germplasm collectively encompassing relatively broad pedigree origins. The majority of the PZA SNPs were developed through the use of a diverse set of 14 maize and 16 teosinte inbreds (Wright et al. 2005) that are expected to result in limited ascertainment bias in maize germplasm. The PHM SNPs were developed from a set of 60 public maize inbreds that have been used in maize breeding in North America and Europe, and include relatively broad pedigree origins encompassing non-stiff stalk, stiff stalk, flint and tropical lines. While these inbreds may not be appropriate for very broad (genus wide) assessments of maize diversity, they are considered appropriate for the analysis of diversity in US and European commercial germplasm. Expected heterozygosity values within each heterotic grouping are provided so that, if needed, researchers can select the most informative markers for specific studies.

The average frequency of polymorphisms detected here, where sequencing covered both the coding region and a portion of the 3′ untranslated region, was one SNP every 43 bases and one indel every 186 bases. This frequency is in the range of values found by others (Tenaillon et al. 2001; Bhattramakki et al. 2002; Ching et al. 2002; Batley et al. 2003; Vroh Bi et al. 2006) with variations being correlated according to whether sequencing focused on the 3′ untranslated region or the less polymorphic coding region. Average nucleotide diversity (θ) was 0.0033, which is lower than has previously been observed in maize. Tenaillon et al. (2001) examined 21 loci in 9 US inbreds and found θ values ranging from 0.0028 to 0.036. Our lower values could be due to the inclusion of a handful of diverse tropical inbreds resulting in some low frequency alleles and subsequently lower nucleotide diversity. Nonetheless, the values we found still represent at least a 4-fold increase over the diversity levels found in humans (Zwick et al. 2000).

The number of SNPs for which an assay could be developed using the generalized design criteria we employed was only 7% of the total number of SNPs discovered, reflecting the high frequency of SNPs in maize that can interfere with marker design. Underlying sequence haplotypes were found to have expected heterozygosity levels twice as high as

SNPs. This higher polymorphism level is on par with multi-allelic markers such as SSRs and RFLPs (Ching et al. 2002; Bhattramakki et al. 2002; Jones et al. 2007). However it should be noted that the SNP haplotypes described here are based on all of the SNPs in a particular sequence, while the number of haplotypes based on SNPs that could be readily interrogated would be considerably less. The full discrimination power of SNP haplotypes may only be realized through the development of SNP haplotype tags that can both meet design criteria and represent the haplotype (Ching et al. 2002), or else through utilizing technologies that allow all of the SNPs in a sequence to be assayed, such as mini-sequencing technologies (eg Pettersson et al. 2003).

Two genetic mapping methods were used; mapping in the IBM2 intermated B73 × Mo17 population and mapping using parent-progeny triplets of Pioneer proprietary inbreds. Some SNPs were also mapped according to their association with an individual accession number, but this method was not consistently reliable; 3/13 SNPs that were assigned to the same accession sequence by BLASTN analysis but that were also independently mapped with genetic methods had locations on different chromosomes. Such inconsistency is most likely due to amplification of unlinked paralogous sequences. In a study by Emrich et al. (2007), one third of the 14 sequences underlying 'Near identical paralogs' (NIPs) were found to be genetically unlinked.

IBM2 map locations determined here showed a high degree of colinearity with the public NAM maps. Out of 282 markers common to the IBM2 and NAM maps, 88.6% (250/282) were collinear, 10.6% (30/282) mapped within 2 cm and 1 marker mapped within 8 cm on the NAM map.

Of the markers incorporated into the Illumina 768 multiplex assay, 591 have map locations on the IBM2 map covering 88% of the genome. These represent a substantial number of SNP markers that can be used for many applications, including for variety identification of North American and European germplasm.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389–3402. doi:10.1093/nar/25.17.3389

Batley J, Mogg R, Edwards D, O'Sullivan H, Edwards KJ (2003) A high-throughput SNuPE assay for genotyping SNPs in flanking regions of *Zea mays* sequence tagged simple sequence repeats. Mol Breed 11:111–120. doi:10.1023/A:1022446021230

Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, 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. doi:10.1023/A:1014841612043

Ching A, Caldwell KS, Jung M, Dolan M, Smith OS, Tingey S, Morgante M, Rafalski AJ (2002) SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. BMC Genet 3:19. doi:10.1186/1471-2156-3-19

Clark A, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of the human genome-wide polymorphism. Genome Res 15:1496–1502. doi:10.1101/gr.4107905

Cone K, McMullen M, Vroh Bi I, Davis G, Yim YS, Gardiner J, Polacco M, Sanchez-Villeda H, Fang Z, Schroeder S, Havermann SA, Bowers JE, Paterson AH, Soderland CA, Engler FW, Wing RA, Coe EH (2002) Genetic, physical and informatic resources for maize: on the road to an integrated map. Plant Physiol 130:1598–1605. doi:10.1104/pp.012245

Emrich SJ, Li L, Wen T-J, Yandeau-Nelson MD, Fu Y, Guo L, Chou H-H, Aluru S, Ashlock DA, Schnable PS (2007) Nearly identical paralogs: implications for maize (*Zea mays* L.) genome evolution. Genetics 175:429–439. doi:10.1534/genetics.106.064006

Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online 1:47–50

Gardiner J, Schroeder SS, Polacco ML, Sanchez-Villeda H, Fang Z, Morgante M, Landewe T, Fengler K, Useche F, Hanafey M, Tingey S, Cou H, Wing R, Soderlund C, Coe EH Jr (2004) Anchoring 9371 maize expressed sequence tagged unigenes to the bacterial artificial chromosome contig map by two-dimentional overgo hybridization. Plant Physiol 134:1317–1326. doi:10.1104/pp.103.034538

Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535

Jones E, Sullivan H, Bhattramakki D, Smith J (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. doi:10.1007/s00122-007-0570-9

Lee MN, Sharopova N, Beavis WD, Grant D, Katt M, Blair D, Hallauer A (2002) Expanding the genetic map of maize with the intermated B73 × Mo17 (IBM) population. Plant Mol Biol 48:453–461. doi:10.1023/A:1014893521186

Pettersson M, Bylund M, Alderborn A (2003) Molecular haplotype determination using allele-specific PCR and pyrosequencing technology. Genomics 82:390–396. doi:10.1016/S0888-7543(03)00177-0

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100. doi:10.1016/S1369-5266(02)00240-6

Rafalski A, Tingey S (2008) SNPs and their use in maize. In: Henry R (ed) Plant Genotyping II: SNP Technology. CAB International, Wallingford, UK, pp.30–43

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Tenaillon MI, Sawkins MC, Long AD, Gaut RL, Doebley JF, Gaut BS (2001) Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. mays L.). Proc Natl Acad Sci USA 98:9161–9166. doi:10.1073/pnas.151244298

Vroh Bi I, McMullen MD, Villeda HS, Schroeder S, Gardiner J, Polacco M, Soderlund C, Wing R, Fang Z, Coe EH (2006) Single nucleotide polymorphisms and insertion-deletions for genetic markers and anchoring the maize fingerprint contig physical map. Crop Sci 46:12–21. doi:10.2135/cropsci2004.0706

Watterson GA (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7:256–276. doi:10.1016/0040-5809(75)90020-9

Wright SI, Vroh Bi I, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, Gaut BS (2005) The effects of artificial selection on the maize genome. Science 308:1310–1314. doi:10.1126/science.1107891

Yu J, Holland JB, McMullen M, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genet 178:539–551. doi:10.1534/genetics.107.074245

Zhao W, Canaran P, Jurkuta R, Fulton T, Glaubitz J, Buckler E, Doebley J, Gaut B, Goodman M, Holland J, Kresovich S, McMullen M, Stein L, Ware D (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res 34:D725–D757. doi:10.1093/nar/gkl196

Zwick ME, Cutler DJ, Chakavarti A (2000) Patterns if genetic variation in Mendelian and complex traits. Annu Rev Genomics Hum Genet 1:387–407. doi:10.1146/annurev.genom.1.1.387