Euphytica (2015) 202:129–139
DOI 10.1007/s10681-014-1238-6

Parental genome contribution in maize DH lines derived from six backcross populations using genotyping by sequencing

Veronica Ogugo · Kassa Semagn · Yoseph Beyene · Steven Runo · Michael Olsen · Marilyn L. Warburton

Received: 20 February 2014 / Accepted: 11 August 2014 / Published online: 24 August 2014

Abstract Molecular characterization of doubled haploid (DH) maize lines and estimation of parental genome contribution (PGC) may be useful for choosing pairs of DH lines for hybrid make up and new pedigree starts. Six BC_1-derived DH populations created by crossing two donor with three recurrent parents were genotyped with 97,190 polymorphic markers with the objectives of: (i) understanding genetic purity, genetic distance and relationship among 417 maize DH lines; (ii) estimating PGC of the DH lines derived from different genetic backgrounds; and (iii) understanding the correlation between donor parent introgression and testcross performance for grain yield and anthesis-silking interval (ASI) under managed drought and optimum environments. The DH lines were 97 % genetically pure, with <2 % heterogeneity; only two DH lines showed heterogeneity >5 %, which is likely to be due to errors during seed multiplication or maintenance. Genetic distance between pairwise comparisons of the 417 DH lines ranged from 0.055 to 0.457; only 0.01 % showed a genetic distance <0.100, indicating large genetic differences among the DH lines. Both populations 1 and 6 showed significantly lower ($p < 0.001$) donor introgression than the other four populations. Donor parent contribution was significantly ($p < 0.001$) higher in the CML444 genetic background than CML395 and CML488. The average donor and recurrent PGC across all 417 DH lines was 31.7 and 64.3 %, respectively. Donor genome introgression was higher than expected in 82 % of the DH lines in the BC_1 generation, possibly due to artificial selection during the DH process, during the development of F_1 or BC_1 seed, or during initial agronomic evaluation of the DH lines. Donor parent introgression up to 32 % showed significant positive correlation with grain yield under drought ($r = 0.312$, $p < 0.001$) and optimum ($r = 0.142$, $p < 0.050$) environments but negative correlation with ASI under drought ($r = -0.276, p < 0.001$). Additional multi-environment phenotype data under managed drought are needed to confirm the correlations reported in this study and to map the specific genomic regions associated with such correlations.

Electronic supplementary material The online version of this article (doi:10.1007/s10681-014-1238-6) contains supplementary material, which is available to authorized users.

V. Ogugo · K. Semagn (✉) · Y. Beyene
International Maize and Wheat Improvement Center (CIMMYT), P. O. Box 1041, Village Market, Nairobi 00621, Kenya
e-mail: k.semagn@cgiar.org

S. Runo
Department of Biochemistry and Biotechnology, Kenyatta University, P. O. Box 43844, Nairobi 0100, Kenya

M. Olsen
International Maize and Wheat Improvement Center (CIMMYT), Apdo. Postal 6-641, 06600 Mexico, D.F., Mexico

M. L. Warburton
United States Department of Agriculture-Agricultural Research Service, Corn Host Plant Resistance Research Unit, Box 9555, Mississippi, MS 39762, USA

Keywords Donor parent · Doubled haploid · Drought · GBS · Introgression · Water-stress

Introduction

Maize (*Zea mays*) is widely grown throughout the world in a range of agro-ecological environments. In sub-Saharan Africa (SSA), maize is a staple food for more than 300 million people and is commonly grown by small-scale and resource poor farmers in rural areas (Shiferaw et al. 2011). Because maize production in SSA is primarily rain-fed, maize yield variability is much higher than in other regions. Several factors, including periodic drought, scarcity of irrigation water, and farmers' inability to use farm inputs contribute to low productivity in the region. Given the unpredictable nature of drought and climate variability over years, breeders must develop improved maize hybrids that are able to withstand drought stress without significant yield penalty under optimal rainfall conditions (Campos et al. 2004; Ribaut and Ragot 2007; Sambatti and Caylor 2007).

The International Maize and Wheat Improvement Center (CIMMYT), in collaboration with Monsanto Company, the African Agricultural Technology Foundation (AATF), and the National Agricultural Research Systems (NARS) of South Africa, Kenya, Uganda, Tanzania and Mozambique, are working together as part of the Water Efficient Maize for Africa (WEMA) project to develop drought tolerant maize for SSA using conventional breeding, marker assisted recurrent selection (MARS) and transgenic technology. WEMA conventional breeding develops new inbred lines by crossing elite lines within heterotic groups followed by sequential self-pollination or the development of doubled haploid (DH) lines. Approximately 1000 maize DH lines were generated by WEMA in 2010 from 10 tropical backcross (BC_1) populations developed by means of in vivo haploid induction (Beyene et al. 2011, 2013). Six of the 10 DH populations were derived by crossing two drought tolerant donor lines extracted from LaPostaSeqC7 (Edmeades et al. 1989) to three CIMMYT inbred lines (CML395, CML444 and CML488) that are well adapted to SSA conditions, particularly to common biotic stresses. The present study focuses on these 6 DH populations. A number of DH-derived hybrids developed from these populations out-yielded commercial checks both under well-watered (optimum) and water-stressed (drought) environments (Beyene et al. 2011, 2013). Superior test-cross performance of some DH lines may be correlated with the proportion of donor parent genome and the parental combination used in the original crosses.

Expectation of parental genome contribution (PGC, the proportion of the genome of an inbred or DH line derived from a specific parent) in populations produced from biparental crosses are easily calculated (Bernardo et al. 1997; Brenner et al. 2012; Frisch and Melchinger 2007; Heckenberger et al. 2006; Melchinger et al. 2010; Prigge et al. 2008). For BC_1-derived lines, expected PGC with Mendelian inheritance is 75 % for the recurrent parent and 25 % for the donor parent. However, recombination during meiosis in the F_1 is random, as is the genetic composition of the daughter cells following migration of the chromatids to each new cell. Therefore, PGC can vary in both amount and specific chromosomal content from each parent. Variation in PGC among BC_1-derived DH lines in the WEMA project has not been characterized. Furthermore, elucidation of the relationship between DH PGC and testcross performance of DH-derived hybrids evaluated both under managed drought and optimum environments may enable more efficient selection.

Estimation of PGC in progeny derived from biparental populations requires genetic data and graphical genotyping software. Advances in next generation technologies have reduced the cost of DNA sequencing to the point that genotyping by sequencing (GBS) can now generate between 0.5 and 1.6 million data points per entry at a genotyping cost ranging from $20 to $38 per sample (http://igd.cornell.edu/index.cfm/page/GBS/GBSpricing.htm). Graphical genotyping (Young and Tanksley 1989) summarizes locus calls into a graphical image of an individual by linkage group or chromosome. This method has been used to identify specific regions of the genome that are positively associated with a desirable trait (Berloo et al. 2001; Foolad et al. 2001; Hayano-Saito et al. 1998; McCouch et al. 1997); to select individuals carrying the positive form of these genomic regions (Severson and Kassner 1995; van Berloo 1999); and to trace the inheritance of specific genomic regions through a pedigree breeding scheme (Ndjiondjop et al. 2008; Semagn et al. 2007).

Table 1 Summary of the 6 DH populations used in the present study

Population code	Donor parent (DP)	Recurrent parent (RP)	Genetic distance between parents[a]	No. of DH lines genotyped	No. of test crosses phenotyped
1	LaPostaSeqC7-F96-1-2-1-1-B–B–B	CML395	0.3641	71	63
2	LaPostaSeqC7-F96-1-2-1-1-B–B–B	CML444	0.3181	85	73
3	LaPostaSeqC7-F96-1-2-1-1-B–B–B	CML488	0.3990	30	28
4	LaPostaSeqC7-F71-1-2-1-2-B–B–B	CML395	0.3668	120	104
5	LaPostaSeqC7-F71-1-2-1-2-B–B–B	CML444	0.3272	86	80
6	LaPostaSeqC7-F71-1-2-1-2-B–B–B	CML488	0.4028	25	25
Total				417	373

[a] Genetic distance is based on 97,190 markers

The objectives of the present study were (i) to understand the genetic purity, genetic distance and relationship among 417 DH lines derived from six BC_1 populations using GBS; (ii) to estimate the variation in PGC among DH lines derived from different genetic backgrounds; and (iii) to understand if the proportion of donor parent introgression is correlated with grain yield and ASI under drought or optimum environments.

Materials and methods

Sample preparation and genotyping

Approximately 1,000 DH lines were developed from 10 BC_1 populations through in vivo haploid induction at the Monsanto facility in Mexico. As the donor parents were drought tolerant but disease susceptible, DH lines were developed from BC_1F_1 to minimize the proportion of donor introgression. Quality control analysis (Semagn et al. 2012) for checking genetic purity of each DH line was conducted by genotyping each line with a total of 165 SNPs by both the Monsanto Company and CIMMYT. Results from QC analysis indicated greater than 5 % genetic heterogeneity (the number of markers that were not homozygous due to mixture of two homozygous genotypes or heterozygosity due to pollen contamination during seed multiplication) for approximately 10 % of the DH lines (Semagn, unpublished). DH lines with greater than 5 % heterogenity were discarded and the remaining DH lines were retained for further characterization. The present study was conducted on 417 DH lines developed from six BC_1 populations created by crossing two drought tolerant but disease susceptible parents (LaPostaSeqC7-F71 and LaPostaSeqC7-F96) with three locally popular and well adapted CIMMYT Maize Lines (CML395, CML444 and CML488) as recurrent parents (Table 1). These populations were selected due to their relatively larger population size and availability of phenotype data. The other four populations had <25 DH lines per population and were not phenotyped so they were excluded in our study.

Leaf samples were harvested from 10 healthy plants per DH line about 3 weeks after sowing at the CIMMYT experimental station in Kiboko, Kenya. The leaves were sampled in perforated Ziploc bags, immediately transferred into a Styrofoam box containing dry ice and transported to the Biosciences eastern and central Africa (BecA) hub in Nairobi. Approximately equal amount of frozen leaf tissue from each of the ten plants per DH line was bulked, cut into pieces with scissors, and transferred into 1.2 mL strip tubes that contained two 4-mm stainless steel grinding balls. In the absence of contamination, DH lines are expected to be genetically pure and could be represented by a single plant genotype data. However, we have frequently observed some level of variability on DH lines due to off-types, stray pollen contamination or seed admixture. Thus, a SNP data of a single plant per entry could be misleading. We therefore extracted DNA for each DH line by bulking about equal amount of leaf tissue from 10 individuals. Samples were ground into fine powder using a tissue grinder (GenoGrinder 2000) at 1,500 strokes per minute for 2 min. Genomic DNA was extracted using a modified version of the Cetyl Trimethyl Ammonium Bromide (CTAB) method of CIMMYT protocol as described elsewhere (Semagn 2014). DNA concentration was measured using the

Quant-iT™ PicoGreen® dsDNA assay kit (Invitrogen™, Paisley, UK) and the Tecan Infinite F200 Pro Plate Reader (Grödig, Austria), and normalized to 50 ng/uL by adding the required volume of 0.1 TE (10 mM Tris–HCl, pH 7.5 and 0.1 mM EDTA, pH 8.0). The quality of the extracted DNA was checked by digesting 250 ng of the genomic DNA from 12 randomly selected samples per 96-well plate with 3.6 units of ApeKI restriction enzyme (New England Biolabs, Boston, USA) at 75 °C for 3 h. Digested DNA samples, along with Lambda DNA digested with Hind III, were run on a 1 % agarose gel containing 0.3 μg/mL GelRed (Biotium Inc., Hayward, USA) and visualized. Fifty uL of the normalized DNA was transferred into a twin.tec® PCR 96-well plate (Eppendorf, Hauppauge, USA) and shipped to the Institute for Genomic Diversity (Cornell University, Ithaca, USA) for genotyping. DNA samples were genotyped using GBS at Cornell University as described by Elshire and colleagues (Elshire et al. 2011).

Phenotyping

Three hundred and seventy three of the 417 DH lines (Table 1) were crossed to a single-cross tester from the opposite heterotic group. Resulting hybrids, along with three commercial and two local checks, were evaluated at three locations (Kenyan Agricultural Research Institute stations in Embu, Kakamega and Kirinyaga Technical Institute) under optimal agronomic conditions, and one managed drought location (CIMMYT station in Kiboko) in Kenya. Trials were planted in two-row plots, 5 m spaced at 0.75 m between rows and 0.25 m between hills using an alpha lattice design, with three replications per location. Testcross performance was evaluated for 10 different traits, but only grain yield and ASI were included in the present study. Details of DH-derived hybrid makeup and testcross evaluation both under drought and optimum conditions have been described elsewhere (Beyene et al. 2011, 2013).

Data analyses

We received imputed GBS data from the Institute for Genomic Diversity (IGD), Cornell University for 940,222 loci (Table 2) per DH line. Since GBS generates a large percentage of un-called genotypes, the missing data was imputed by IGD using an algorithm that searched for the closest neighbor in small SNP windows across maize database available at IGD (Romay et al. 2013). Genotype data was filtered using a minor allele frequency (MAF) of 0.05 using TASSEL version 4.1.7 software (Bradbury et al. 2007). This filtering resulted to 97,190 polymorphic markers (10.3 % of the initial loci) for further analyses (Table 2). The proportion of heterogeneity and missing data after imputation were computed for each DH line. Identity-by-state (IBS) genetic distance was calculated between each pair of lines using TASSEL and used for hierarchical clustering analysis using the neighbor joining method implemented in DARwin version 5.0.158 (http://darwin.cirad.fr/Home.php). Principal coordinate analysis (PCoA) was performed on the genetic distance matrix using DARwin, and the first two principal components were plotted for visual examination of the clustering pattern of the DH lines.

For each population, a dataset consisting of a subset of 97,190 polymorphic markers was created for graphical genotyping. To improve marker inheritance calculations and marker distribution on each chromosome, additional filtering was done to exclude markers that were (a) missing in one or both parents, (b) monomorphic between the two parents, (c) heterogeneous in one or both parents, and/or (d) physically linked (<50 kb). The final number of markers retained for graphical genotyping varied from 3,724 to 6,665 (Table 2), and the average was 4,846 markers per population. Input files for graphical genotyping were prepared using ParentChecker version 1 (Hu et al. 2012). The proportions of the donor and recurrent parent genomes were estimated using GGT software version 2 (van Berloo 1999).

For phenotypic data, repeatability for single environments and broad-sense heritability in the combined analyses across environments were computed using SAS as described in a previous study (Semagn et al. 2013). Frequency distributions and Pearson correlations between parental genome contribution and the phenotypic traits were computed using Minitab version 14.

Results

Genetic characterization

The MAF and missing data after imputation for the 97,190 polymorphic markers varied from 0.051 to

Table 2 Summary of the number of markers used for genotyping and analysis

Chromosome	Total markers	Polymorphic markers	Number of markers used for graphical genotyping					
			Population 1	Population 2	Population 3	Population 4	Population 5	Population 6
1	143,834	13,643	648	1,037	532	767	722	586
2	113,467	13,365	672	835	429	649	550	510
3	105,637	11,163	572	789	435	618	479	493
4	98,895	7,850	484	695	375	567	537	467
5	105,177	10,936	440	599	323	542	504	393
6	77,136	9,681	394	592	295	465	420	338
7	80,124	8,467	385	566	577	420	379	343
8	80,008	8,586	354	633	263	404	448	347
9	69,892	5,752	392	433	222	403	299	245
10	66,052	7,747	392	486	273	371	364	322
Total	940,222	97,190	4,733	6,665	3,724	5,206	4,702	4,044

0.500 (averaging 0.119) and from 0 to 19.4 % (averaging 4.6 %), respectively (data not shown). The number of polymorphic markers per chromosome varied from 5,752 on chromosome 9 to 13,643 on chromosome 1, with a mean of 9,719 (Table 2). Heterogeneity and missing data per DH line varied from 0 to 7.5 % (averaging 0.4 %) and from 2.5 to 10.0 % (averaging 4.6 %), respectively (Supplementary material 1). Approximately 92 % of the DH lines had heterogeneity $\leq$1 %; only two DH lines showed heterogeneity >5 %. Due to the bulk genotyping, heterogeneity would be contributed either by heterozygosity or a combination of two homozygous SNPs. For the 2 DH lines with >5 % heterogeneity, the source of such unexpected level of heterogeneity can be determined by re-genotyping DNA extracted from single plants, with each line represented by 5 to 10 individual plants. The bulking method significantly reduces the time and cost required in processing 5-10 individual plants from each of the 417 DH lines (2,085–4,170 plants in total) to only 10–20 plants from the 2 DH lines that showed high proportion of heterogeneity. Genetic distance between pairwise comparisons of the 417 DH lines ranged from 0.055 to 0.457 (Fig. 1), with an overall average of 0.286. Most genetic distances between lines from different populations ranged from 0.301 to 0.400 (Fig. 1) and within populations from 0.100 to 0.300. Only 13 out of 86,736 pairwise comparisons (0.01 %) had a genetic distance <0.100, clearly indicating lack of redundant lines even among DH lines developed from the same cross. The neighbor joining tree generated from the distance matrix grouped the 417 DH lines into three major groups based on recurrent parent, although a subgroup formed in one of the clusters based on donor parent (Supplementary material 2). The first three principal components (PCs) from principal coordinate analysis explained 58.6 % of the total SNP variation among samples. A plot of PC1 (37.7 %) and PC2 (17.4 %) also formed 3 major groups based on recurrent parent; however, no sub-grouping based on donor parent was seen (Supplementary material 2).

Parental genome contribution

The average recurrent parent genome across populations varied from 61.6 % in population 3 to 66.2 % in population 1 (Table 3), with an overall average of 64.1 %. Population 1 had significantly higher ($p = 0.03$) recurrent parent genome than both populations 2 and 3; the other 3 populations had intermediate level of recurrent parent genome. The average donor genome introgression across the 6 populations varied from 27.7 % in population 6 to 33.9 % in population 5, and the overall average was 31.7 % (Table 3; Fig. 2). The proportions of LaPostaSeqC7-F96 genome in population 1 and LaPostaSeqC7-F71 genome in population 6 were significantly lower ($p < 0.001$) than the other four populations (Table 3). Of the 3 recurrent parents, CML444 had significantly higher ($p < 0.001$) donor introgression (33.6 %) than both CML395 (30.3 %) and CML488 (30.4 %).

The proportion of donor genome introgression among the individual DH lines was highly variable,

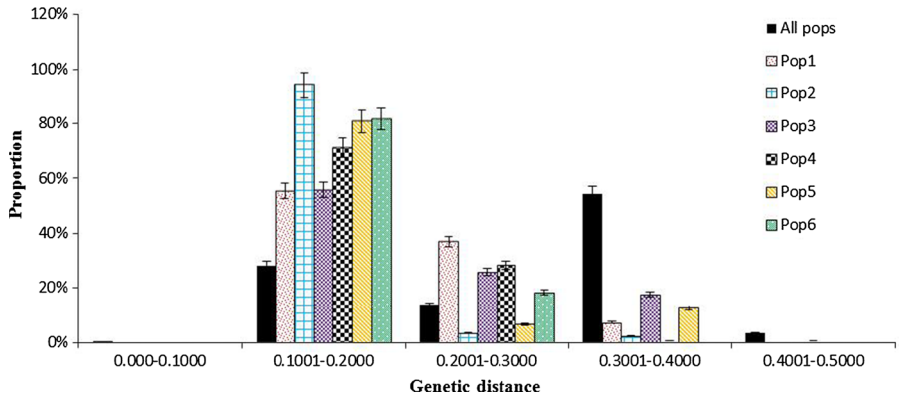

Fig. 1 Frequency distribution of pairwise genetic distance among DH lines within and among populations

Table 3 Summary of the recurrent and donor parents genome contribution across the 6 DH populations

Donor parent (DP)	Recurrent parent (RP)	No. of DH lines	Population	RP (%)	DP (%)	Heterozygote (%)	Missing (%)
LaPostaSeqC7-F96-1-2-1-1-B–B–B	CML395	71	1	66.2	28.2	0.8	4.8
LaPostaSeqC7-F96-1-2-1-1-B–B–B	CML444	85	2	62.6	33.5	0.7	3.3
LaPostaSeqC7-F96-1-2-1-1-B–B–B	CML488	30	3	61.6	32.7	0.5	5.1
LaPostaSeqC7-F71-1-2-1-2-B–B–B	CML395	120	4	65.1	31.6	0.6	2.8
LaPostaSeqC7-F71-1-2-1-2-B–B–B	CML444	86	5	63.8	33.9	0.6	1.8
LaPostaSeqC7-F71-1-2-1-2-B–B–B	CML488	25	6	65.0	27.7	1.0	6.4

ranging from 16.3 to 69.2 % (Supplementary material 3). About 18 % of the DH lines had donor genome introgression lower than the expected 25 % introgression at BC_1 generation, while the remaining 82 % had >25 % donor introgression. Nearly 45 % of the DH lines had donor introgression higher than the 31.7 % average donor introgression across all 6 populations. The frequency distribution of donor introgression was not statistically ($p = 0.342$) different between the two donor parents (Supplementary material 4). Donor parent introgression among populations showed significant ($p < 0.05$) differences within chromosomes (Fig. 3), but which chromosome had the lowest introgression varied between populations. For example, population 6 had the lowest introgression on chromosomes 2, 5, 6, 7 and 8, while populations 3 and 4 had the lowest donor introgression on chromosome 1.

Correlation between parental genome contribution and testcross performance

Heritability for grain yield under drought, optimum and combined analyses of both water regimes were 0.530, 0.445 and 0.430, respectively. For ASI, heritability under drought, optimum and combined analyses were 0.711, 0.494 and 0.545, respectively. Donor genome introgression across all 373 DH lines showed significant positive correlation with grain yield ($r = 0.155$, $p < 0.001$) and negative correlation with ASI ($r = -0.177$, $p < 0.001$) under drought but it was not significantly correlated with either grain yield or ASI under optimum environments (Table 4). Correlation analyses between parental genome contribution and grain yield and ASI were also conducted by dividing the 373 DH lines into 2 groups based on the proportion of donor introgression (≤32 % and >32 % introgression, 32 % being the average overall donor introgression, Fig. 2). For DH lines with ≤32 % donor introgression, donor parent genome showed significant positive correlation with grain yield under drought ($r = 0.312$, $p < 0.001$) and optimum ($r = 0.142$, $p < 0.050$), and negative correlation with ASI ($r = -0.276$, $p < 0.001$) under drought. Neither trait showed significant correlation with donor parent genome when the introgression was higher than 32 %.

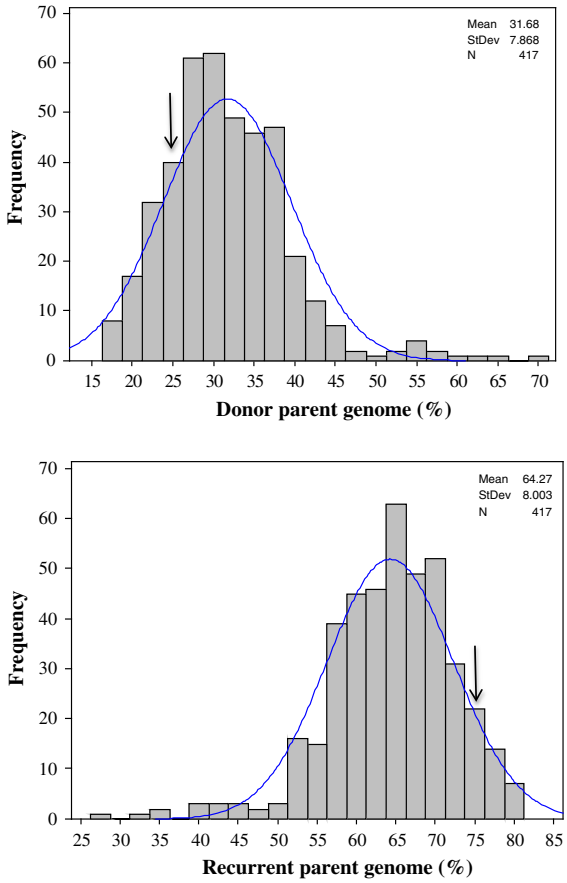

Fig. 2 Frequency distribution of donor parent (*top*) and recurrent parent (*bottom*) genome contribution among the 417 DH lines. The expected donor and recurrent parent genome contribution at BC_1 generation are indicated in arrows

Discussion

Genetic characterization

Maize breeding revolves around the development of improved inbred lines (either using conventional breeding methods or DH technology) and identifying the best parental combinations for creating hybrids that are higher yielding than their parents (Duvick 2001). Maize inbred lines are primarily developed by crossing elite lines within heterotic groups, while hybrids are produced by crossing inbred lines that belong to different heterotic groups. The genetic uniformity of hybrids depends on the genetic purity of the parents used in making the cross. In the present study, all DH lines generated for the WEMA project were initially genotyped for QC analysis and DH lines with over 5 % heterogeneity were discarded. However, even following QC, two DH lines still showed >5 % heterogeneity, which may have been caused by pollen contamination during seed multiplication, off-types from prior varieties grown in the field, and seed admixtures during harvesting and handling. Other studies have also reported substantial proportion of heterozygosity (heterogeneity) within some DH lines (Heckenberger et al. 2002; Murigneux et al. 1993).

Pairwise comparisons of the genetic distance matrix of the 417 DH lines showed the presence of ample genetic variation (a prerequisite for gain via selection) among the DH lines both between and within populations. In other studies (Beyene et al. 2011, 2013), small subsets of hybrids derived using some of the DH lines as parents were evaluated across multiple managed drought and optimum environments. Combined analyses across drought and optimum environments showed that some of the top hybrids derived using the DH lines as parents produced up to 2.2 t/ha under optimum and up to 1.4 t/ha under water-limiting conditions higher grain yield than the mean of the commercial checks (Beyene et al. 2011; 2013). Such superior phenotypic performance of the DH-derived hybrids may be due to the observed high genetic distance among the DH lines.

Parental genome contributions

In a conventional backcrossing program, BC_1-derived progenies are expected to contain 25 % of the donor and 75 % of the recurrent parent genome, but actual percentages often vary considerably (Brenner et al. 2012), as they did in this study (Supplementary material 3). Such deviation in donor parent genome introgression from the expectation may be due to one or more of the following factors: (i) the DH process (induction, genome doubling, and seed increase); (ii) artificial selection during the initial agronomic evaluation of the DH lines; (iii) genetic drift during selfing, and (iv) chance fluctuation. The DH process is stringent and produces a very small proportion of viable haploid kernels and fertile DH lines. In one study, only about 6 % of 4,400 putative haploid kernels produced viable and fertile DH lines (Wilde et al. 2010). During the process of developing an improved progeny line through pedigree selection or DH technology, breeders also usually select the most vigorous BC plants. These plants are generally more

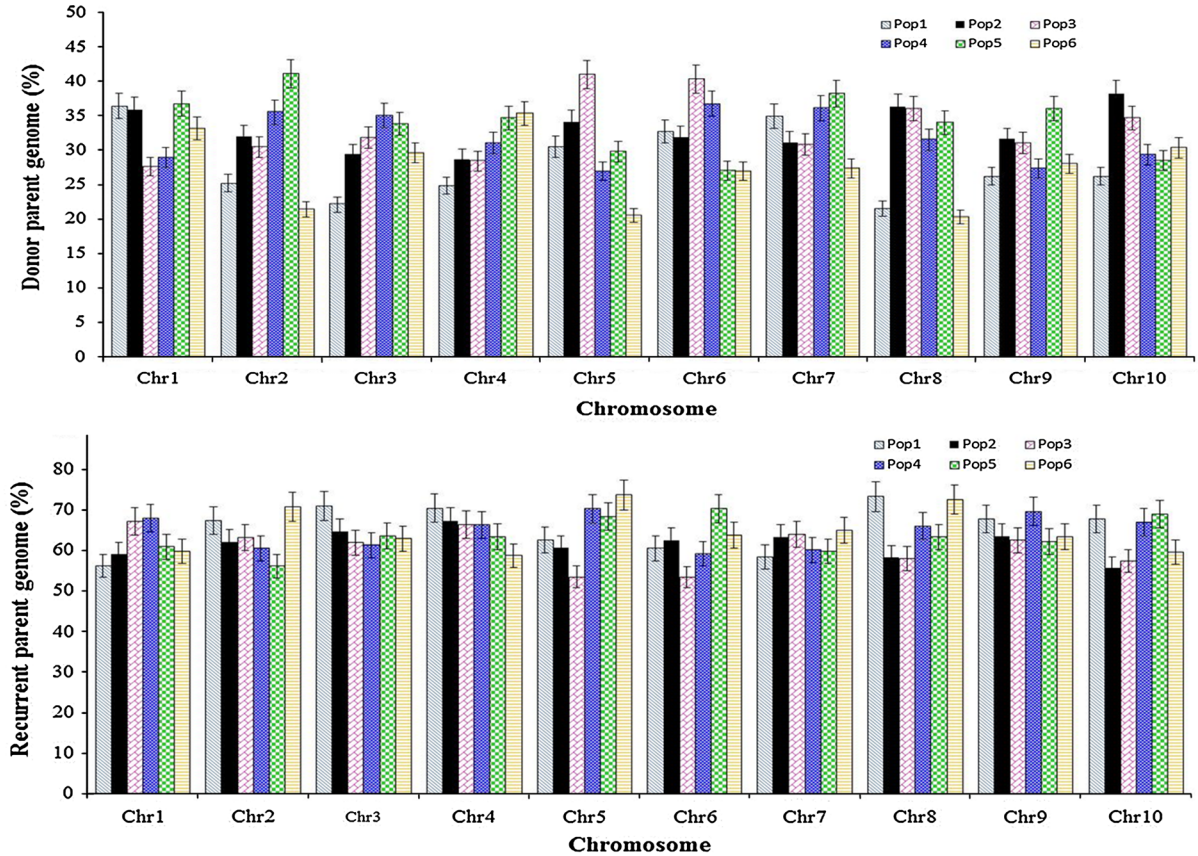

Fig. 3 The proportion of donor (*top*) and recurrent (*bottom*) parent genome contribution in each chromosome for 6 DH populations. Populations 1–3 and 4–6 were derived using LaPostaSeqC7-F96 and LaPostaSeqC7-F71 as donor parents, respectively

Table 4 Correlation between parental genome contribution and grain yield and anthesis-silking interval (ASI)

Trait	Category	Sample size	Parent	Drought	Optimum
Grain yield	All DH lines	373	Donor	0.155**	0.051
			Recurrent	−0.094	0.071
	DH lines with ≤ 32.0 % donor introgression	210	Donor	0.312**	0.142*
			Recurrent	−0.167*	0.069
	DH lines with >32 % donor introgression	163	Donor	0.015	0.094
			Recurrent	0.060	−0.004
Anthesis-silking interval	All DH lines	373	Donor	−0.177**	−0.100
			Recurrent	0.162**	0.109*
	DH lines with ≤ 32.0 % donor introgression	210	Donor	−0.276**	−0.074
			Recurrent	0.214**	0.087
	DH lines with > 32 % donor introgression	163	Donor	−0.096	−0.106
			Recurrent	0.082	0.178*

* $p < 0.050$; ** $p < 0.001$

heterozygous than expected for randomly drawn BC plants and possess a higher proportion of the donor genome, which shifts their distribution in the direction of F_2-derived progeny (Heckenberger et al. 2005). Phenotypic selection among the DH lines for target management without specific selection in favor of the recurrent parent may also cause higher donor and lower recurrent parent genome across the DH populations. The DH lines in this study were evaluated both under managed drought and optimum environments, and only lines that showed good performance under drought without yield penalty under optimum environments were selected.

Results from correlation analyses between donor PGC and agronomic traits showed the presence of low but significant correlation (both positive and negative) with grain yield and ASI, respectively. The proportion of the donor parent to be incorporated to the adapted germplasm is one of the major concerns in backcross introgression programs. In order to determine the optimal proportion of donor parent genome for phenotypic performance of the DH-derived hybrids, we computed the correlation between grain yield and ASI within two categories of donor introgression. The correlation between donor PGC and grain yield or ASI under drought was nearly double when only DH lines with $\leq$32 % introgression were used in the analysis than when all DH lines were analyzed together (Table 4). There was no correlation between donor PGC and grain yield or ASI under drought when donor introgression exceeded 32 %. This may be caused by simultaneous transfer of undesirable genes flanking beneficial genes from the donor parent (linkage drag). The average grain yield under optimal water conditions of hybrids derived from DH lines with $\leq$32 % or with >32 % donor introgression was the same. These results suggest that when using drought tolerant donor lines to improve performance of SSA-adapted lines under water-limiting conditions, the optimal range of donor PGC should approach, but not exceed, 32 %. One possible application of these findings would be to cull lines with sub-optimal donor PGC prior to phenotypic evaluation. Because DH-derived hybrids were evaluated for phenotypic performance only at one managed drought screening environment, additional data may be needed to confirm the correlations reported in this study and also to map the specific genomic regions associated with the two traits.

Genotyping cost

A few hundred SNPs uniformly distributed across all maize chromosomes may be sufficient for estimating parental genome contribution among lines derived from biparental crosses. In the present study, the total number of polymorphic SNPs across the 417 DH lines was 97,190 but only an average of 4,846 SNPs per population were used for computing parental genome contributions. The total GBS genotyping cost per sample, including informatics was $33, which translates to $13,959 for all DH lines and their parents. If the same number of samples were genotyped with the chip-based 1536 SNPs using Illumina GoldenGate platform (http://www.illumina.com), the total genotyping cost would vary between $25,380 and $50,760 ($60–120 per sample depending on the sample size). If they were genotyped with high quality pre-selected 250 SNPs using the Kompetitive Allele Specific PCR (KASP) at the LGC Genomics (http://www.lgcgenomics.com), the total genotyping cost would be $15,228 ($36 per sample) (Semagn et al. 2014). The genotyping cost per data point would therefore be $0.007 US for GBS (assuming only an average of 4846 SNPs was used for analysis), $0.044 to $0.078 US for GoldenGate and $0.144 US for KASP. Therefore, GBS is at least 6 and 20 times cheaper than GoldenGate and KASP platforms, respectively, with a marker density at least 3 and 19 times higher than the two platforms. GBS genotyping cost per data point decreases to $0.00034 when all the 97,190 polymorphic SNPs were considered for analysis. Although GBS provides low-cost and high-density genotype information, it does generate a substantial amount of missing data and a non-uniform distribution of sequence reads (Beissinger et al. 2013).

Conclusion

Results from the present study reveal the presence of high genetic distance among nearly all pairwise comparisons of 417 DH lines studied. Significantly higher donor parent genomic contribution compared to theoretical expectations was also observed. Low to moderate but significant correlation between grain yield and ASI under drought stress with donor parent genome introgression up to 32 % suggests that detrimental linkage drag may occur when the

proportion of donor introgression exceeds 32 %, and that culling sub-optimal donor PGC DH lines prior to phenotyping may improve efficiency of breeding for drought tolerance. As the DH-derived hybrids were evaluated for phenotypic performance only at one managed drought site, additional phenotype data may be needed to confirm the correlations reported in this study and also map the specific genomic regions associated with such correlations.

Acknowledgments DH lines used in this study and their phenotypic data were generated as part of the WEMA project, funded by the Bill & Melinda Gates Foundation. GBS data were generated as part of the Basic Research to Enabling Agricultural Development (BREAD) project, funded by the US National Science Foundation. We are grateful to CIMMYT Field Technicians at the different stations in Kenya for the phenotypic evaluation, and Monsanto Company for developing the DH populations.

References

Beissinger TM, Hirsch CN, Sekhon RS, Foerster JM, Johnson JM, Muttoni G, Vaillancourt B, Buell CR, Kaeppler SM, de Leon N (2013) Marker Density and read-depth for genotyping populations using genotyping-by-sequencing. Genetics 193:1073–1081

Bernardo R, Murigneux A, Maisonneuve JP, Johnsson C, Karaman Z (1997) RFLP-based estimates of parental contribution to F2- and BC1-derived maize inbreds. Theor Appl Genet 94:652–656

Beyene Y, Mugo S, Pillay K, Tefera T, Ajanga S, Njoka S, Karaya H, Gakunga J (2011) Testcross performance of doubled haploid maize lines derived from tropical adapted backcross populations. Maydica 56:351–358

Beyene Y, Mugo S, Semagn K, Asea G, Trevisan W, Tarekegne A, Tefera T, Gethi J, Kiula B, Gakunga J, Karaya H, Chavangi A (2013) Genetic distance among doubled haploid maize lines and their testcross performance under drought stress and non-stress conditions. Euphytica 192:379–392

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635

Brenner EA, Blanco M, Gardner C, Lübberstedt T (2012) Genotypic and phenotypic characterization of isogenic doubled haploid exotic introgression lines in maize. Mol Breed 30:1001–1016

Campos H, Cooper M, Habben JE, Edmeades GO, Schussler JR (2004) Improving drought tolerance in maize: a view from industry. Field Crops Res 90:19–34

Duvick DN (2001) Biotechnology in the 1930s: the development of hybrid maize. Nat Rev Genet 2:69–74

Edmeades GO, Bolanos J, Lafitte HR, Rajaram S, Pfeiffer W, Fischer RA (1989) Traditional approaches to breeding for drought resistance in cereals. In: Proceedings of a symposium held in Cairo, Egypt, 28–30 November, pp 27–52

Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379

Foolad MR, Zhang LP, Lin GY (2001) Identification and validation of QTLs for salt tolerance during vegetative growth in tomato by selective genotyping. Genome 44:444–454

Frisch M, Melchinger AE (2007) Variance of the parental genome contribution to inbred lines derived from biparental crosses. Genetics 176:477–488

Hayano-Saito Y, Tsuji T, Fujii K, Saito K, Iwasaki M, Saito A (1998) Localization of the rice stripe disease resistance gene, Stv-bi, by graphical genotyping and linkage analyses with molecular markers. Theor Appl Genet 96:1044–1049

Heckenberger M, Bohn M, Ziegle JS, Joe LK, Hauser JD, Hutton M, Melchinger AE (2002) Variation of DNA fingerprints among accessions within maize inbred lines and implications for identification of essentially derived varieties. I. Genetic and technical sources of variation in SSR data. Mol Breed 10:181–191

Heckenberger M, Bohn M, Melchinger AE (2005) Identification of essentially derived varieties obtained from biparental crosses of homozygous lines: I. Simple sequence repeat data from maize inbreds. Crop Sci 45:1120–1131

Heckenberger M, Muminovic J, Voort JR, Peleman J, Bohn M, Melchinger AE (2006) Identification of essentially derived varieties obtained from biparental crosses of homozygous lines. III. AFLP data from maize inbreds and comparison with SSR data. Mol Breed 17:111–125

Hu Z, Ehlers J, Roberts P, Close T, Lucas M, Wanamaker S, Xu S (2012) ParentChecker: a computer program for automated inference of missing parental genotype calls and linkage phase correction. BMC Genet 13:9

McCouch SR, Chen X, Panaud O, Temnykh S, Xu Y, Cho Y, Huang N, Ishii T, Blair M (1997) Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol 35:89–99

Melchinger AE, Dhillon BS, Mi X (2010) Variation of the parental genome contribution in segregating populations derived from biparental crosses and its relationship with heterosis of their Design III progenies. Theor Appl Genet 120:311–319

Murigneux A, Barloy D, Leroy P, Beckert M (1993) Molecular and morphological evaluation of doubled haploid lines in maize. 1. Homogeneity within DH lines. Theor Appl Genet 86:837–842

Ndjiondjop MN, Semagn K, Sie M, Cissoko M, Fatondji B, Jones M (2008) Molecular profiling of interspecific lowland rice populations derived from IR64 (Oryza sativa) and Tog5681 (*Oryza glaberrima*). Afr J Biotechnol 7:4219–4229

Prigge V, Maurer HP, Mackill DJ, Melchinger AE, Frisch M (2008) Comparison of the observed with the simulated distributions of the parental genome contribution in two marker-assisted backcross programs in rice. Theor Appl Genet 116:739–744

Ribaut JM, Ragot M (2007) Marker-assisted selection to improve drought adaptation in maize: the backcross approach, perspectives, limitations, and alternatives. J Exp Bot 58:351–360

Romay M, Millard M, Glaubitz J, Peiffer J, Swarts K, Casstevens T, Elshire R, Acharya C, Mitchell S, Flint-Garcia S, McMullen M, Holland J, Buckler E, Gardner C (2013) Comprehensive genotyping of the USA national maize inbred seed bank. Genome Biol 14:R55

Rv Berloo, Aalbers H, Werkman A, Niks RE (2001) Resistance QTL confirmed through development of QTL-NILs for barley leaf rust resistance. Mol Breed 8:187–195

Sambatti JBM, Caylor KK (2007) When is breeding for drought tolerance optimal if drought is random? New Phytol 175:70–80

Semagn K (2014) Leaf tissue sampling and DNA extraction protocols. In: Besse P (ed) Molecular plant taxonomy: methods and protocols. Human Press, New York, pp 53–67

Semagn K, Ndjiondjop MN, Lorieux M, Cissoko M, Jones M, McCouch S (2007) Molecular profiling of an interspecific rice population derived from a cross between WAB 56-104 (*Oryza sativa*) and CG 14 (*Oryza glaberrima*). Afr J Biotechnol 6:2014–2022

Semagn K, Beyene Y, Makumbi D, Mugo S, Prasanna BM, Magorokosho C, Atlin G (2012) Quality control genotyping for assessment of genetic identity and purity in diverse tropical maize inbred lines. Theor Appl Genet 125:1487–1501

Semagn K, Beyene Y, Warburton M, Tarekegne A, Mugo S, Meisel B, Sehabiague P, Prasanna B (2013) Meta-analyses of QTL for grain yield and anthesis silking interval in 18 maize populations evaluated under water-stressed and well-watered environments. BMC Genom 14:313

Semagn K, Babu R, Hearne S, Olsen M (2014) Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. Mol Breed 33:1–14

Severson DW, Kassner VA (1995) Analysis of mosquito genome structure using graphical genotyping. Insect Mol Biol 4:279–286

Shiferaw B, Prasanna BM, Hellin J, Bänziger M (2011) Crops that feed the world 6. Past successes and future challenges to the role played by maize in global food security. Food Secur 3:307–327

van Berloo R (1999) Computer note. GGT: software for the display of graphical genotypes. J Hered 90:328–329

Wilde K, Burger H, Prigge V, Presterl T, Schmidt W, Ouzunova M, Geiger HH (2010) Testcross performance of doubled-haploid lines developed from European flint maize landraces. Plant Breed 129:181–185

Young ND, Tanksley SD (1989) Restriction fragment length polymorphism maps and the concept of graphical genotypes. Theor Appl Genet 77:95–101