Genotyping-by-sequencing and microarrays are complementary for detecting quantitative trait loci by tagging different haplotypes in association studies

Sandra Silvia Negro[1,3], Emilie Millet[2,4], Delphine Madur[1], Cyril Bauland[1], Valérie Combes[1], Claude Welcker[2], François Tardieu[2], Alain Charcosset[1], Stéphane Dimitri Nicolas[1]

Affiliations :

[1] GQE – Le Moulon, INRA, Univ. Paris-Sud, CNRS, AgroParisTech, Université Paris-Saclay, 91190 Gif-sur-Yvette, France.

[2] Laboratoire d'Ecophysiologie des Plantes sous Stress Environnementaux (LEPSE), UMR759, INRA-SupAgro, 34060 Montpellier, France.

[3] Present address: GIGA-R Medical Genomics, University of Liège, Belgium.

[4] Present address: Biometris, Department of Plant Science, Wageningen University & Research, 6700 AA Wageningen, The Netherlands.

E-mail addresses:

snegro@uliege.be; emilie.millet@wur.nl; delphine.madur@inra.fr; cyril.bauland@inra.fr; valerie.combes@inra.fr; claude.welcker@inra.fr; francois.tardieu@inra.fr; alain.charcosset@inra.fr; stephane.nicolas@inra.fr

Corresponding author:

stephane.nicolas@inra.fr

Abstract

Background: Single Nucleotide Polymorphism (SNP) array and re-sequencing technologies have different properties (*e.g.* calling rate, minor allele frequency profile) and drawback (*e.g.* ascertainment bias), which lead us to study the complementarity and consequences of using them separately or combined in diversity analyses and Genome-Wide Association Studies (GWAS). We performed GWAS on three traits (grain yield, plant height and male flowering time) measured in 22 environments on a panel of 247 diverse dent maize inbred lines using three genotyping technologies (Genotyping-By-Sequencing, Illumina Infinium 50K and Affymetrix Axiom 600K arrays).

Results: The effects of ascertainment bias of both arrays were negligible for deciphering global genetic trends of diversity in this panel and for estimating relatedness. We developed an original approach based on linkage disequilibrium (LD) extent in order to determine whether SNPs significantly associated with a trait and that are physically linked should be considered as a single QTL or several independent QTLs. Using this approach, we showed that the combination of the three technologies, which have different SNP distribution and density, allowed us to detect more Quantitative Trait Loci (QTLs, gain in power) and potentially refine the localization of the causal polymorphisms (gain in position).

Conclusions: Conceptually different technologies are complementary for detecting QTLs by tagging different haplotypes in association studies. Considering LD, marker density and the combination of different technologies (arrays and re-sequencing), the genotypic data presently available were most likely enough to well represent polymorphisms in the centromeric regions, whereas using more markers would be beneficial for telomeric regions.

Keywords: GWAS, linkage disequilibrium, genome coverage, maize, high-throughput genotyping technologies.

Background

Understanding the genetic bases of complex traits involved in the adaptation to biotic and abiotic stress in plants is a pressing concern, with world-wide drought due to climate change as a major source of human food and agriculture threats. Recent progress in next generation sequencing and genotyping array technologies contribute to a better understanding of the genetic basis of quantitative trait variation by performing Genome-Wide Association Studies (GWAS) on large diversity panels [1]. Single Nucleotide Polymorphism (SNP)-based techniques became the most commonly used genotyping methods for GWAS because SNPs are cheap, numerous, codominant and can be automatically analysed with SNP-arrays or produced by genotyping-by-sequencing (GBS), or sequencing [2-4]. The decreasing cost of genotyping technologies have led to an exponential increase in the number of markers used for the GWAS in association panels, thereby raising the question of computation time to perform the association tests. Computational issues were addressed by using either approximate methods by avoiding re-estimating variance component for each SNP [5] or exact methods using mathematical tools for sparing time in matrix inversion [6, 7]. It is noteworthy that using approximate computation in GWAS can produce inaccurate p-values when the SNP effect size is large or/and when the sample structure is strong [8].

Several causes may impact the power of Quantitative Trait Locus (QTL, locus involved in quantitative trait variation) detection in GWAS. Highly diverse panels have in general undergone multiple historical recombinations, leading to a low extent of linkage disequilibrium (LD). However, these panels can present different average and local patterns of LD [9-11]. A high marker density and a proper distribution of SNPs are therefore essential to capture causal polymorphisms. Furthermore, minor allele frequencies (MAF), population stratification and cryptic relatedness are three other important parameters affecting power and

false positive detection [12, 14]. These last two factors are substantial in several cultivated species such as maize [15] and grapevine [16], and their impact on LD can be statistically evaluated [17]. Population structure and kinship can be estimated using molecular markers [18-21] and can be modelled to efficiently detect marker-trait associations due to linkage only [12, 22, 23]. These advances have largely increased the power and effectiveness of linear mixed models that can now efficiently account for population structure and relatedness in GWAS [12, 8].

In maize, an Illumina Infinium HD 50,000 SNP-array (50K array) was developed by Ganal *et al.* [3] and has been used extensively for diversity and association studies [24, 25]. For example, GWAS were conducted to unravel the genetic architecture of phenology, yield component traits and to identify several flowering time QTLs linked to adaptation of tropical maize to temperate climate [26, 27]. With the same array, Rincent *et al.* [11] showed that LD occurs over a longer distance in a dent than in a flint panel, with appreciable effects on the power of QTL detection. Comparison of LD extent between association panels suggests that genotyping with 50K markers causes a limited power of GWAS in many panels due to the low LD extent and the correlation between allelic values at some SNPs due to the kinship and population structure [14, 27]. Therefore, higher marker densities are desirable because the maize genome size is large (2.4 Gb), the level of diversity is high, and LD extent is low (more than one substitution per hundred nucleotides) [28]. As a consequence, an Affymetrix Axiom 600,000 SNP-array (600K array) was developed and used in association genetics [29, 30] and detection of selective sweep [4]. Another possibility is whole genome sequencing, but this is currently unpractical for large genomes such as maize because of the associated cost. Hence, a Genotyping-By-Sequencing (GBS) procedure has been developed [2] that targets low-copy

genomic regions by using cheap restriction enzymes. Genotyping-by-sequencing has been successfully used in maize for genomic prediction [31]. Romay *et al.* [32] and Gouesnard *et al.* [33] highlighted the interest of the GBS for (i) deciphering and comparing the genetic diversity of the inbred lines in seedbanks and (ii) identifying QTLs by GWAS for kernel colour, sweet corn and flowering time. To our knowledge, the respective interests of DNA arrays and GBS for diversity analyses and GWAS have never been compared in plants.

The main drawback of the DNA arrays is that they do not allow to discover new SNPs. This possibly leads to some ascertainment bias in diversity analysis when the SNPs selected for building arrays come from (i) the sequencing of a set of individuals who did not well represent the diversity explored in the studied panel, (ii) a subset of SNPs that skews the allelic frequency profile towards the intermediate frequencies [27, 34]. Ascertainment bias can compromise the ability of the SNP arrays to reveal an exact view of the genetic diversity [34]. Genotyping-by-sequencing can overcome ascertainment bias since it is based on sequencing and therefore allows the discovery of alleles in the diversity panel analysed. It can be generalized to any species at a low cost providing that numerous individuals have been sequenced in order to build a representative library of short haplotypes to call SNPs [35]. Non-repetitive regions of genomes can be targeted with two- to three-fold higher efficiency, thereby considerably reducing the computationally challenging problems associated with alignment in species with high repeat content. However, GBS may have a low-coverage leading to a high missing data rate (65% in both studies; [32, 33]) and heterozygote under-calling, depending on genome size and structure, and on the number of samples combined in the flow-cell. Furthermore, GBS requires the establishment of demanding bioinformatic pipelines and imputation algorithms [36]. Pipelines have been developed to call SNP

119 genotypes from raw GBS sequence data and to impute the missing data from a haplotype
120 library [35, 36].

121 Here, we investigated the impact of using GBS and DNA arrays on the quality of the
122 genotyping data, together with the biological properties of data generated by these
123 technologies, and the potential complementarity of these approaches. In particular, we
124 analyzed the impact of increasing the marker density and using different genotyping
125 technologies (sequencing *vs* array) on (i) the estimates of relatedness and population structure,
126 (ii) the detection of QTLs (power. To address these issues, we performed a GWAS based on
127 genotypic datasets obtained using either GBS or DNA arrays with low (50k) or high (600k)
128 densities on a diversity panel of maize hybrids obtained from a cross of dent lines with a
129 common flint tester. Three traits were considered, namely grain yield, plant height and male
130 flowering time (day to anthesis), measured in 22 different environments (sites × years ×
131 treatments) over Europe. We developed an original approach based on LD extent in order to
132 determine whether SNPs significantly associated with a trait should be considered as a single
133 QTL or several independent QTLs.

134

135 **Results**

Combining Tassel and Beagle imputations improved the genotyping quality for GBS

136 We estimated the genotyping and imputation concordance of the GBS based on common
137 markers with the 50K or 600K arrays (Additional file 1: Figure S1 and Table S1). After SNP
138 calling from reads using AllZeaGBSv2.7 database (direct reads, GBS_1, Additional file 1:
139 Figure S1), the call rate was 33.81% on the common SNPs with the 50K array, *vs* 37% for the

whole GBS dataset. The genotyping concordance rate was 98.88% (Additional file 1: Table S1). After imputation using *TASSEL* by Cornell Institute (GBS_2), the concordance rate was 96.04% on the common markers with the 50K array and 11.91% of missing data remained for the whole GBS dataset. In GBS_3, all missing data were imputed by *Beagle* but yielded a lower concordance rate (92.14% and 91.58% for the 50K and the 600K arrays). In an attempt to increase the concordance rate of the genotyping while removing missing data, we tested two additional methods, namely GBS_4 where the missing data and heterozygotes of Cornell imputed data (GBS_2) were replaced by Beagle imputation, and GBS_5 where Cornell homozygous genotypes (GBS_2) were completed by imputations from GBS_3 (Additional file 1: Figure S1 and Table S1). GBS_5 displayed a slightly better concordance rate than GBS_2 (96.25% vs 96.04%) and predicted heterozygotes with a higher quality than GBS_4. GBS_5 was therefore used for all genetic analyses and named GBS hereafter.

GBS displayed more rare alleles and lower call rate than microarrays

The SNP call rate was higher for the arrays (average values of 96% and >99% for the 50K and 600K arrays, respectively), than for the GBS (37% for the direct reads). The MAF distribution differed between the technologies (Figure 1): while the use of arrays resulted in a near-uniform distribution, GBS resulted in an excess of rare alleles with a L-shaped distribution (22% of SNPs with MAF < 0.05 for the GBS *versus* 6% and 9% for the 50K and 600K, respectively). This is not surprising since the 50K array was based on 27 sequenced lines for SNPs discovery [3], the 600K array was based on 30 lines for [4], whereas GBS was based on 31,978 lines, thereby leading to higher discovery of rare alleles. Consistent with MAF distribution, the average gene diversity (*He*) was lower for GBS (0.27) than for arrays (0.35 and 0.34 for the 50K and 600K arrays, respectively). The distribution of SNP

heterozygosity was similar for the three technologies, with a mean of 0.80%, 0.89% and 0.22 % for the 50K and 600K arrays and GBS, respectively. The heterozygosity of inbred lines was highly correlated between technologies with large coefficients of Spearman correlation: $r_{50K-600K} = 0.90$, $r_{50K-GBS} = 0.76$, $r_{600K-GBS} = 0.83$. The distribution of the SNPs along the genome was denser in the telomeres for the GBS and in the peri-centromeric regions for the 600K array, whereas the 50K array exhibited a more uniform distribution (top graph in Figure 2 and in Additional file 2: Figure S2).

Population structure and relatedness were consistent between the three technologies

We used the ADMIXTURE software to analyse the genetic structure within the studied panel based on SNPs from the three technologies, by using two to ten groups. Based on a K-fold cross-validation, the clustering in four genetic groups ($N_Q = 4$) was identified as the best one in the datasets resulting from the three technologies. Considering a threshold of 0.5 (ancestral fraction), the assignation to the four groups was identical except for a few admixed inbred lines (Additional file 3: Figure S3). Based on the 50K, the four groups were constituted by (i) 39 lines in the Non Stiff Stalk (Iodent) family traced by PH207, (ii) 46 lines in the Lancaster family traced by Mo17 and Oh43, and (iii) 55 lines in the stiff stalk families traced by B73 and (iv) 107 lines that did not fit into the three primary heterotic groups, such as W117 and F7057. This organization appeared consistent with the organization of breeding programs into heterotic groups, generally related to few key founder lines.

We compared two estimators of relatedness between inbred lines, IBS (Identity-By-State) and *K_Freq* (Identity-By-Descend), calculated per technology. For IBS, pairs of individuals were

on average more related using the GBS data than those from arrays (Additional file 3: Table S2). Relatedness estimated with the two arrays were highly correlated: r = 0.95 and 0.98 for IBS and *K_Freq*, respectively (Figure 3 and Additional file 3: Figure S4b). The differences between the kinships estimated from the three technologies were reduced if the excess of rare alleles in the GBS was removed (Additional file 3: Figure S4c).

We further carried out diversity analyses by performing Principal Coordinate Analyses (PCoA) on IBD (*K_Freq*, weights by allelic frequency) estimated from the three technologies. The two first PCoA axes explained 12.9%, 15.6% and 16.3% of the variability for the GBS, 50K and 600K arrays, respectively. The same pattern was observed regardless of the technology with the first axis separating the Stiff Stalk from the Iodent lines and the second axis separating the Lancaster from the Stiff Stalk and Iodent lines (see illustration with the 50K kinship, Additionnal File 3: Figure S5). Key founders lines of the three heterotic groups (Iodent: PH207, Stiff Stalk: B73, Lancaster: Mo17) were found at extreme positions along the axes, which was consistent with the admixture groups previously described.

Long distance linkage disequilibrium was removed by taking into account population structure or relatedness

In order to evaluate the effect of kinship and the genetic structure on linkage disequilibrium (LD), we studied genome-wide LD between 29,257 PANZEA markers from the 50K array within and between chromosomes before and after taking into account the kinship (*K_Freq* estimated from the 50K array), structure (Number of groups = 4) or both (Additional file 4: Figure S6). Whereas inter-chromosomal LD was only partially removed when the genetic structure was taken into account, it was mostly removed when either the kinship or both

kinship and structure were considered (Additional file 4: Figure S6b and c). Accordingly, long distance intra-chromosomal LD was almost totally removed for all chromosomes by accounting for the kinship, structure or both. Interestingly, some pairs of loci located on different chromosomes or very distant on a same chromosome remained in high LD despite correction for genetic structure and kinship (Additional file 4: Figure S6). This can be explained either by genome assembly errors, by chromosomal rearrangements such as translocations or by strong epistatic interactions. Linkage disequilibrium decreased with genetic or physical distance, Additional file 4: Figure S7). The majority of pairs of loci with high LD (r^2K>0.4) in spite of long physical distance (>30Mbp), were close genetically (<3cM), notably on chromosome 3, 5, 7 and to a lesser extent 9 and 10. These loci were located in centromeric and peri-centromeric regions that displayed low recombination rate, suggesting that this pattern was due to variation of recombination rate along the chromosome. Only very few pairs of loci in high LD were genetically distant (>5cM) but physically close (<2Mbp). Linkage disequilibrium (r^2K and r^2KS) was negligible beyond 1 cM since 99% of LD values were less than 0.12 in this case. Note that some unplaced SNPs remained in LD after taking into account the kinship and structure with some SNPs with known positions on chromosome 1, 3 and 4 (Additionnal File 4: Figure S6). Therefore, LD measurement corrected by the kinship can help to map unplaced SNPs.

Linkage disequilibrium strongly differed between and within chromosomes

We combined the three technologies together to calculate the r^2K for all pairs of SNPs, which were genetically distant by less than 1 cM. For any chromosome region, LD extent in terms of genetic and physical distance showed a limited variation over the 100 sets of 500,000 loci pairs (cf. Material). This suggests that the estimation of LD extent did not strongly depend on

our set of loci. LD extent varied significantly between chromosomes for both high recombinogenic (>0.5 cM/Mbp) and low recombinogenic regions (<0.5 cM/Mbp, Table 1). Chromosome 1 had the highest LD extent in high recombination regions (0.062 ±0.007 cM) and chromosome 9 the highest LD extent in low recombinogenic regions (898.6±21.7 kbp) (Table 1). Linkage disequilibrium extent relative to genetic and physical distances was highly and positively correlated in high recombinogenic regions (r = 0.86), whereas it was not in low recombinogenic regions (r = -0.64).

Table 1: Variation of LD extent, and percentage of genome covered.

		Chromosome										Whole Genome
		1	2	3	4	5	6	7	8	9	10	
Physical Size (kbp)		301,354	237,069	232,140	241,474	217,873	169,174	176,765	175,794	156,751	150,189	2,058,583
Genetic Size (cM)		268	211	188	150	205	129	148	182	145	139	1766
Physical LD extent (kbp) in Low Recombination regions (<0.5 cM / Mbp)		306	491	846	808	658	418	547	497	899	815	629
Genetic LD Extent (cM) in High Recombination regions (>0.5 cM / Mbp)		0.062	0.027	0.033	0.022	0.031	0.019	0.012	0.038	0.023	0.019	0.029
Percent of physical genome covered	50K	81%	72%	76%	77%	74%	67%	71%	73%	71%	71%	74%
	600K	98%	88%	91%	89%	90%	84%	81%	90%	87%	84%	89%
	GBS	92%	81%	84%	83%	83%	77%	77%	81%	79%	76%	82%
	Combined	98%	90%	92%	90%	91%	87%	83%	92%	88%	85%	90%
Percent of genetic map covered	50K	72%	41%	44%	38%	41%	32%	24%	46%	32%	27%	42%
	600K	96%	71%	76%	68%	72%	62%	47%	78%	63%	53%	71%
	GBS	86%	58%	61%	53%	61%	48%	37%	63%	48%	40%	58%
	Combined	97%	74%	78%	72%	74%	65%	51%	81%	66%	57%	74%

Genetic and Physical LD extent were obtained by adjusting Hill and Weir model's on 100 different sets of 500,000 loci randomly sampled in high and low recombination regions, respectively. The value represented the average across these 100 sets. The percentage of genome coverage was estimated using markers with MAF > 5% and E(r^2k) = 0.1, for each technology and for the three technologies combined (GBS+600K+50K).

The effective population size (*Ne*) estimated from the Hill and Weir's model [37] using genetic distance varied from 7.9±0.04 (Chromosome 1) to 41.2±0.12 (Chromosome 7) in high recombinogenic regions. Noteworthy, the same approach lead to unrealistic values in low

recombinogenic regions (from 961 on Chromosome 6 to more than 1 million for chromosome 2 and 10), thereby confirming that the use of genetic distance is not well suited to model local LD in low recombinogenic regions.

Finally, we studied the variation of local LD extent along each chromosome by adjusting the Hill and Weir's model against genetic distance within a sliding windows of 1cM (Additional file 4: Figure S8). After removing intervals that did not reach our criteria (Absence of model convergence, effective population size > 247, low number of loci), the 3,205 remaining intervals (90%) showed a high variation for genetic LD extent along each chromosome, with LD extent varying from 0.019 (*Ne* = 246) to 0.997 cM (*Ne* = 0.06) (Additional file 4: Figure S8).

Large differences in genome coverage between technologies

We estimated the percentage of the genome that was covered by LD windows around SNPs, calculated by using either physical or genetic distances (Table 1). We observed a strong difference in coverage between the three technologies at both genome-wide and chromosome scale, as illustrated in Figure 2 on chromosome 3 (Table 1, and Additional file 2: Figure S2). For a LD extent of r^2K = 0.1, 74%, 82% and 89% of the physical map, and 42%, 58% and 71% of the genetic map were covered by the 50K array, the GBS and the 600K array, respectively (Table 1). For the combined data (50K + 600K + GBS), the coverage strongly varied between chromosomes, ranging from 83% (chromosome 7) to 98% (chromosome 1) of the physical map, and from 51% (chromosome 7) to 97% (chromosome 1) of the genetic map (Table 1). For the physical map, increasing the LD extent threshold to r^2K=0.4 reduced the genome coverage from 89% to 49% for 600K, 82% to 28% for GBS, 74% to 20% for 50K and 90% to 52% for the combined data. Increasing the MAF threshold reduced slightly the

genome coverage, with smaller reduction for the physical map than genetic map. Surprisingly, increasing the SNP number by combining the markers from the arrays and GBS did not strongly increase the genome coverage as compared to the 600K, regardless of the threshold for LD extent (Figure 2 and Additional file 2: Figure S2).

We observed a strong variation of genome coverage along each chromosome with contrasted patterns in low and high recombinogenic regions (Figure 2 and Additional file 2: Figure S2). While low recombinogenic regions were totally covered with all the technologies (except for few intervals using the 50K array), the genome coverage in high recombinogenic regions varied depending on both technology and SNP distribution. 47% of the 2Mbp intervals in high recombination regions were better covered by the 600K array than the GBS against only 1%, which were better covered by GBS than 600K.

Number of QTLs detected using genome-wide association studies increases with markers density

We observed a strong variation in the number of SNP significantly associated with the three traits across the 22 environments (Table 2). The mean number of significant SNPs per environment and trait was 3.7, 44.7, 17.9 and 62.4 for the 50K, 600K, GBS and the three technologies combined, respectively (Table 3). Considering the p-value threshold used, 28, 303 and 204 false positives were expected among the 243, 2,953 and 1,182 associations detected for 50K, 600K and GBS, respectively. False discovery rate appeared therefore higher for GBS (17.2%) than for DNA arrays (11.5% and 10.2% for 50K and 600K, respectively). It could be explained by the higher genotyping error rate of GBS due to imputation and/or by its higher number of makers with a lower MAF. Both reduce the power of GBS compared to

DNA arrays and therefore lead to a higher false discovery rate. Proportionally to the SNP number, 50K and 600K arrays resulted in 1.5- and 1.7-fold more associated SNPs per situation (environment × trait) than GBS (p-value<$2x10^{-6}$, Table 3). This difference between arrays and GBS was higher for grain yield (GY) and plant height (plantHT) than for male flowering time (DTA, Table3).

Table 2: Number of significant SNPs per environment, per technology and for the combined technologies.

	Flowering Time					Plant Height					Grain Yield				
Environment	Heritability	50K	600K	GBS	Combined	Heritability	50K	600K	GBS	Combined	Heritability	50K	600K	GBS	Combined
Cam12R	0.36	0	3	1	4	0.52	23	209	88	289	0.35	21	286	102	381
Cam12W	0.60	1	18	13	31	0.43	22	270	72	339	0.52	41	525	167	684
Cam13R	0.45	0	9	7	16	0.15	0	2	3	5	0.18	0	1	3	4
Cra12R	0.64	1	40	18	59	0.26	0	6	3	9	0.23	3	57	45	103
Cra12W	0.69	3	25	19	43	0.18	10	69	16	83	0.54	12	98	53	150
Deb12R	0.58	1	14	16	30	0.25	0	1	0	1	0.56	2	14	0	16
Deb12W	0.73	0	25	38	61	0.41	0	6	7	7	0.47	0	6	2	8
Deb13R	0.60	1	17	5	23	0.08	0	33	9	42	0.35	1	22	15	37
Gai12R	0.62	8	80	24	104	0.15	1	47	41	89	0.31	0	23	8	31
Gai12W	0.66	5	42	15	59	0.40	0	1	3	4	0.56	3	71	14	85
Gai13R	0.58	0	24	8	31	0.56	0	6	6	11	0.66	0	4	5	9
Gai13W	0.78	1	45	9	54	0.40	0	1	3	4	0.81	2	7	1	9
Kar12R	0.71	4	30	21	52	0.26	0	4	3	7	0.73	0	5	6	11
Kar12W	0.77	8	60	10	73	0.22	1	10	4	14	0.54	2	19	11	29
Kar13R	0.66	3	65	11	77	0.27	0	4	2	6	0.92	4	37	24	62
Kar13W	0.81	0	17	12	29	0.26	0	2	7	9	0.67	4	12	6	19
Mur13R	0.85	3	48	19	68	0.26	7	61	7	68	0.84	14	90	28	116
Mur13W	0.8	0	11	8	19	0.33	3	4	2	9	0.74	10	80	25	104
Ner12R	0.70	7	23	18	45	0.22	0	7	3	10	0.53	1	10	6	16
Ner12W	0.80	1	80	30	107	0.28	0	2	2	4	0.60	1	8	6	15
Ner13R	0.77	3	60	26	88	0.22	1	25	13	38	0.35	0	13	7	20
Ner13W	0.81	2	23	17	42	0.27	0	8	5	13	0.76	2	28	4	32
Average		2.4	34.5	15.7	50.6		3.1	35.4	13.6	48.4		5.6	64.4	24.5	87.9
Median		1	25	15.5	47.5		0	6	4.5	9.5		2	20.5	7.5	30

The average, median and standard deviation (SD) per environment are calculated for each trait (male Flowering Time, Plant Height, Grain Yield).

Table 3: Comparison of associated SNPs and QTLs detected between traits and three technologies.

		Significant SNPs				QTLs			
Technology		50K	600K	GBS	Combined	50K	600K	GBS	Combined
Marker Nb		42046	459191	308929	810580	42046	459191	308929	810580
Total Nb	DTA	52	759	345	1115	20	130	133	226
	plantHT	68	778	299	1061	16	96	90	160
	GY	123	1416	538	1941	33	166	120	238
	Per trait	81	984	394	1372	23	131	114	208
Average per envir.	DTA	2.4	34.5	15.7	50.7	0.9	5.9	6.0	10.3
	plantHT	3.1	35.4	13.6	48.2	0.7	4.4	4.1	7.3
	GY	5.6	64.4	24.5	88.2	1.5	7.5	5.5	10.8
	Per trait	3.7	44.7	17.9	62.4	1.0	5.9	5.2	9.5
Average per SNP tested	DTA	5.70E-5	7.50E-5	5.10E-5	2.10E-5	2.20E-5	1.30E-5	2.00E-5	1.30E-5
	plantHT	7.40E-5	7.70E-5	4.40E-5	2.00E-5	1.70E-5	9.50E-6	1.30E-5	9.00E-6
	GY	1.30E-4	1.40E-4	7.90E-5	3.60E-5	3.60E-5	1.60E-5	1.80E-5	1.30E-5
	Per trait	8.80E-5	9.70E-5	5.80E-5	7.70E-5	2.50E-5	1.30E-5	1.70E-5	3.50E-5

QTLs were obtained by grouping associated SNPs with overlapping LD windows (*LD_win*) for the three traits (DTA: male flowering time; PlantHT: plant height; GY: grain yield). *Marker Nb*: is the number of markers tested in GWAS. *Total number*: is the sum of associated SNPs or QTLs across environments. *Average per envir.*: is the average number of QTLs obtained in 22 environments for three traits (66 trait-environments combinations). *Average per SNP tested*: is the number of associated SNPs or QTLs detected divided by the number of SNP tested.

We used two approaches based on LD for grouping significant SNPs: (i) considering that all SNPs with overlapping LD windows for $r^2K=0.1$ belong to the same QTL (*LD_win*) and (ii) grouping significant SNPs that are adjacent on the physical map and are in LD ($r^2K > 0.5$, *LD_adj*). The QTLs defined by using the two approaches were globally consistent since significant SNPs within QTLs were in high LD whereas SNPs from different adjacent QTLs were not (Additional file 6: Figure S9-LD-Adjacent and Additional file 7: Figure S9-LD-Windows). *LD_adj* detected more QTLs than *LD_win* for flowering time (242 vs 226), plant height (240 vs 160) and grain yield (433 vs 237). The number of QTLs detected with the *LD_adj* approach increased strongly when the LD threshold was set above 0.5. Differences in QTL groupings between the two methods were observed for specific LD and recombination patterns. This occurred for instance on chromosome 6 for grain yield (Additional file 6:

Figure S9-LD-Adjacent and Additional file 7: Figure S9-LD-Windows). Within this region, the recombination rate was low and the LD pattern between associated SNPs was complex. While *LD_adj* splitted several SNPs in high LD into different QTLs (for instance QTL 232, 235, 237, 249), *LD_win* grouped together associated SNPs that are genetically close but displayed a low LD (Additional file 6: Figure S9-LD-Adjacent and Additional file 7: Figure S9-LD-Windows). Reciprocally, for flowering time, we observed different cases where *LD_win* separated distant SNPs in high LD into different QTLs whereas *LD_adj* grouped them (QTL 25 and 26, 51 to 53, 95 to 97, 208 and 209, 218 and 219). As these differences were specific to complex LD and recombination patterns, we used the *LD_win* approach for the rest of the analyses.

Although a large difference in number of associated SNPs was observed between 600K and GBS, little difference was observed between QTL number after grouping SNPs (Table 2, Table 3). The mean number of QTLs was indeed 1.0, 5.9 and 5.2 and 9.5 for the 50K, 600K, GBS, and the three technologies combined, respectively (Table 3). Note that the number of QTLs continued to increase with marker density when SNPs from GBS, 50K and 600K were combined (Figure 4). The number of SNPs associated with each QTL varied according to the technology (on average 3.7, 7.6, 3.4 and 6.6 significant SNPs for the 50K, 600K, GBS, and the combined technologies, respectively). The total number of QTLs detected over all environments by using the 600K array and GBS was close for flowering time (130 *vs* 133) and plant height (96 *vs* 90). It was 1.4-fold higher for the 600K than GBS for grain yield (166 *vs* 120).

The 600K and GBS were highly complementary for association mapping

The 600K and GBS technologies were highly complementary to detect QTLs for the three traits: 78%, 76% and 71% of the QTLs of flowering time, plant height and grain yield, respectively, were specifically detected by 600K or GBS (Figure 5). On the contrary, 50K displayed very few specific QTLs. While only 9 out 69 QTLs from the 50K array were not detected when the 600K array was used, 39 QTLs detected using the 50K array were not detected when using GBS. When we combined the GBS and 600K markers, 7% of their common QTLs had $-log_{10}(Pval)$ increased by 2 and 21% by 1 potentially indicating a gain in accuracy of the position of the causal polymorphism (Additional file 8: Table S3).

This complementarity between GBS and 600K is well exemplified with two strong association peaks for flowering time on chromosome 1 (QTL32) and 3 (QTL95) detected in several environments (Additional file 8: Table S3 and Figure 6a). In order to better understand the origin of the complementarity between GBS and 600K technologies for GWAS, we scrutinized the LD between SNPs and the haplotypes within these two QTLs (Figure 6b and c, and Additional file 9: Figure S10 for other examples). For example, QTL95 showed a gain in power. It was only identified by the 600K array although the region included numerous SNPs from GBS close to the associated peak. None of these SNPs was in high LD with the most associated marker of the QTL95 (Figure 6b). Another example is QTL32, which was detected by 1 to 10 GBS markers in 9 environments with *–log(p-value)* ranging from 5 to 7.6, whereas it was detected by only two 600K markers in one environment (Ner12W) with *–log(p-value)* slightly above the significance threshold (Additional file 8: Table S3 and Figure 6b).

Haplotype analyses showed that the SNPs from the GBS within QTL95 were not able to discriminate all haplotypes (Figure 6c). In QTL95, using the 600K markers allowed one to

368 discriminate the three main haplotypes (H1, H2, H3), whereas using the GBS markers did not
369 allow discrimination of H3 against H1 + H2. As H1 contributed to an earlier flowering time
370 than H2 or H3, associations appeared more significant for the 600K than for GBS (Figure 6c).
371 In QTL32, the use of GBS markers allowed identifying late individuals that mostly displayed
372 H1, H2 and H3 haplotypes, against early individuals that mostly displayed H4 and H5
373 haplotypes (Figure 6c). The gain of power for GBS markers as compared to 600K markers for
374 QTL32 originated from the ability to discriminate late individuals (black alleles) from early
375 individuals (red alleles) within H4 haplotypes (Figure 6c).

Stability, pleiotropy and distribution of QTL detected across environments

376 After combining the three technologies, we identified 226, 160, 238 QTLs for the flowering
377 time, plant height and grain yield, respectively (Table 3 and Additional file 8: Table S3). We
378 highlighted 23 QTLs with the strongest effects on flowering time, plant height and grain yield
379 ($-log_{10}(Pval) \geq 8$, Table 4). The strongest association corresponded to the QTL95 for
380 flowering time ($-log_{10}(p\text{-}value) = 10.03$) on chromosome 3 (158,943,646 – 159,005,990 bp),
381 the QTL135 for GY ($-log_{10}(p\text{-}value) = 18.7$) on chromosome 6 (12,258,527 – 29,438,316 bp)
382 and QTL78 on chromosome 6 (12,258,527 – 20,758,095 bp) for plant height ($-log_{10}(p\text{-}value)$
383 = 17.31). The QTL95 for flowering time trait was the most stable QTLs across environments
384 since it was detected in 19 environments (Additional file 8: Table S3). Moreover, this QTL
385 showed a colocalization with QTL74 for grain yield in 5 environments and QTL30 for plant
386 height in 1 environment suggesting a pleiotropic effect. More globally, 472 QTLs appeared
387 trait-specific whereas 70 QTLs overlapped between at least two traits (6,3%, 5.2% and 3.0%
388 for GY and plantHT, GY and DTA, and DTA and plantHT, respectively) suggesting that
389 some QTLs are pleiotropic (Additional file 10: Figure S11). This is not surprising since

average corresponding correlations within environments for these traits were moderate (0.47, 0.54 and 0.45, respectively). Only 0.7% overlapped between the three traits (Additional file 10: Figure S11). Twenty percent of QTLs were detected in at least two environments and 9% in at least three environments (Additional file 10: Figure S12 and Table S4). We observed no significant differences of stability between the three traits (*p-value* = 0.2). However, 6 out 7 most stable QTLs (Number of environments >5) were found for flowering time and a higher proportion of QTLs were specific for plant height than grain yield and flowering time (85% vs 77% for both flowering time and grain yield, *p-value* = 0.09, *p-value* = 0.2), respectively. We observed that QTLs that displayed a significant effect in more than one environment had larger effects and *-log(p-value)* values than those significant in a single environment. This difference in *-log(p-value)* values was stronger for grain yield and plant height than flowering time.

Table 4: Summary of the main QTLs ($-log_{10}(Pval) \geq 8$) identified for the three traits.

Trait	QTL	Chr	LowerLimit	UpperLimit	Pos	R2	Effect	Log	MinorAll	MajorAll	MAF	EnvMax	NbDiffEnv
DTA	95	3	158943646	159005990	158974594	0.15	1.27	10.03	G	C	0.41	Ner13R	19
plantHT	21	2	129971437	130912039	130441738	0.15	-6.74	9.21	A	G	0.14	Cam12W	3
plantHT	71	6	6593785	6636807	6614012	0.13	-4.76	8.39	G	A	0.18	Cam12R	2
plantHT	72	6	6793841	6837747	6807230	0.14	-4.87	8.77	T	G	0.18	Cam12R	2
plantHT	78	6	12258527	20758095	20330595	0.27	-8.99	17.31	C	T	0.26	Cam12W	4
plantHT	79	6	21037721	23951687	22905376	0.19	-5.45	11.42	T	G	0.31	Cam12R	3
plantHT	80	6	24184017	26606537	25317825	0.13	-4.41	8.17	T	C	0.2	Cam12R	2
plantHT	81	6	26695327	28659766	28130108	0.16	-5.14	9.84	C	G	0.44	Cam12R	2
plantHT	94	6	101463249	101501936	101482646	0.14	-5.98	8.22	T	A	0.17	Cam12W	2
plantHT	110	8	12767198	12798330	12782777	0.19	-7.67	12.44	C	T	0.22	Cam12W	3
GY	65	3	140505559	144210207	141621777	0.12	0.42	8.13	A	C	0.27	Gai12W	5
GY	85	3	186994852	187057772	187028970	0.12	-0.48	8.49	A	C	0.28	Kar12W	1
GY	120	6	5131927	5177694	5155708	0.12	-0.58	8.24	C	T	0.42	Cam12W	2
GY	122	6	5623945	5659803	5638516	0.12	-0.57	8.11	T	G	0.23	Cam12W	2
GY	124	6	5855407	5887383	5871000	0.12	-0.56	8.4	T	G	0.42	Cam12W	2
GY	127	6	6593785	6636807	6612654	0.16	-0.65	10.44	C	A	0.32	Cam12W	2
GY	128	6	6793841	6837747	6807462	0.12	-0.54	8.41	A	C	0.26	Cam12W	2
GY	129	6	6878877	6930838	6890199	0.12	0.63	8.06	T	A	0.48	Cam12W	3
GY	130	6	7027497	7088575	7046773	0.13	0.62	8.71	A	G	0.4	Cam12W	3
GY	131	6	7113662	7200479	7159714	0.12	0.59	8.24	T	C	0.39	Cam12W	3
GY	135	6	12258527	29438316	18528943	0.28	-0.78	18.7	G	C	0.31	Cam12W	6
GY	147	6	101463249	101501936	101482646	0.22	-0.65	15.04	T	A	0.17	Cam12W	5
GY	173	8	12767198	12798330	12782777	0.17	-0.61	11.8	C	T	0.22	Cam12W	4

"LowerLimit" and "UpperLimit" are the lower and upper physical limits for each QTL, followed by the physical position (*Pos*, bp), proportion of the variance explained (R^2), the effect of the major allele (*Effect*) as outputted by FastLMM, $-log_{10}$(***Pval***) (*Log*), the minor and major alleles and the minor allele frequency (*MAF*) of the most significant SNP within the QTL. The environment for which the most associated QTL was observed (EnvMax) and the number of different environments (NbDiffEnv) that detected the QTL are shown. Note that QTLs 71-72 for the plant height and QTLs 129-130 for the grain yield are genetically close (<1cM) and display high mean LD (r^2K>0.5). Hence, QTLs 71-72 and 129-130 can potentially be merged.

The distribution of QTLs was not homogeneous along the genome since 82%, 77% and 79% of flowering time, plant height and grain yield QTLs, respectively, were located in the high recombinogenic regions, whereas they represented 46% of the physical genome (Additional file 10: Table S5 and Figure S13). The QTLs were more stable (≥ 2 environments) in low than in high recombinogenic regions (12.8% vs 5.8%, *p-value* = 0.03).

Discussion

GBS required massive imputation but displayed similar global trends than DNA arrays for genetic diversity organization

In order to reduce genotyping cost, GBS is most often performed at low depth leading to a high proportion of missing data, thereby requiring imputation in order to perform GWAS. Imputation can produce genotyping errors that can cause false associations and introduce bias in diversity analysis [33]. We evaluated the quality of genotyping and imputation obtained by different approaches, taking the 50K or 600K arrays as references. The best imputation method that yielded a fully genotyped matrix with a low error rate for the prediction of both heterozygotes and homozygotes was the approach merging the homozygous genotypes from Tassel and the imputation of Beagle for the other data (GBS_5 in Additional file 1: Table S1). The quality of imputation was high with 96% of allelic values consistent with those of the 50K and 600K arrays. This level of concordance is identical than in a study of USA national maize inbred seed bank by Romay *et al.* [32]. It is higher than in a diversity study of European flint maize collection (93%) by Gouesnard *et al.* [33], which was more distant from the reference AllZeaGBSv2.7 database than for the panel presented here.

The ascertainment bias of arrays due to the limited number of lines used for SNP discovery was reinforced by counter-selection of rare alleles during the design process of DNA arrays [3, 4]. For GBS, the polymorphism database to call polymorphisms included thousands of diverse lines [35]. In our study, we used AllZeaGBSv2.7 database. After a first step of GBS imputation (GBS_2), missing data dropped to 11.9% *i.e.* only slightly more than in Romay *et al.* (10%) [34]. This confirms that the polymorphism database (AllZeaGBSv2.7) covered

adequately the genetic diversity of our genetic material.

Although, we observed differences of allelic frequency spectrum between GBS and DNA arrays, these technologies revealed similar trends in the organization of population structure and relatedness (Figure 1, Additional file 3: Figure S3 and S4 and Table S2) suggesting no strong ascertainment bias for deciphering global genetic structure trends in the panel. However, although highly correlated, level of relatedness differed between GBS and DNA arrays, especially when the lines were less related as showed by the deviation (to the left) of the linear regression from the bisector (Figure 3).

The extent of linkage disequilibrium strongly varied along and between chromosomes

Linkage disequilibrium extent in high recombinogenic regions varied to a large extent among chromosomes, ranging from 0.012 to 0.062 cM. Similar variation of genetic LD extent between maize chromosomes has been previously observed by Rincent et al. [14], but their classification of chromosomes was different from ours. This difference could be explained by the fact that we analyzed specifically high and low recombination regions. According to Hill and Weir Model [37], the physical LD extent in a genomic region increased when the local recombination rate decreased. As a consequence, chromosome 1 and 9 had the lowest and highest physical LD extent and displayed the highest and one of the lowest recombination rate in pericentromeric regions, respectively (0.26 vs 0.11 cM / Mbp, Table 1 and Additional file 10: Table S5). Unexpectedly, the genetic LD extent also correlated negatively with the recombination rate. It suggested that chromosomes with a low recombination rate also display a low effective population size. Background selection for deleterious alleles could explain this pattern since it reduced the genetic diversity in low recombinogenic regions [38, 39]. Finally,

462 we observed a strong variation of the LD extent along each chromosome (Additional file 4:
463 Figure S8). As we used a consensus genetic map [40] that represents well the recombination
464 within our population, it suggested, according to Hill and Weir's model, that the number of
465 ancestors contributing to genetic diversity varied strongly along the chromosomes. This likely
466 reflects the selection of genomic regions for adaptation to environment or agronomic traits
467 [38], that leads to a differential contribution of ancestors according to their allelic effects.
468 Ancestors with strong favorable allele(s) in a genomic region may lead ultimately to large
469 identical by descent genomic segments [41].

SNPs were clustered into QTL highlighting interesting genomic regions

470 In previous GWAS, the closest associated SNPs were grouped into QTLs according to
471 either a fixed physical distance [1] or a fixed genetic distance [30, 42]. These approaches
472 suffer of two drawbacks. First, the physical LD extent can vary strongly along chromosomes
473 according to the variation of recombination rate (Additional file 2: Figure S2). Second, the
474 genetic LD extent depends both on panel composition and the position along the genome
475 (Table 1). These approaches may therefore strongly overestimate or underestimate the number
476 of QTLs. To address both issues Cormier *et al.* [43] proposed to group associated SNPs by
477 using a genetic window based on the genetic LD extent estimated by Hill and Weir model in
478 the genomic regions around the associated peaks [37]. In our study, we improved this last
479 approach (*LD_win):*

480 - First, we used r^2K that corrected r^2 for kinship rather than the classical r^2 since r^2K
481 reflected the LD addressed in our GWAS mixed models to map QTL [17].

482 - Second, we took advantage of the availability of both physical and genetic maps of
483 maize to project the genetic LD extent on the physical map. This physical window was useful

484 to retrieve the annotation from B73 reference genome, decipher local haplotype diversity
485 (Figure 6) and estimate physical genome coverage (Table 1, Figure 2).
486 - Third, we considered an average LD extent estimated separately in the high and low
487 recombinogenic genomic regions. This average was estimated by using several large random
488 sets of pairs of loci in these regions rather than the local LD extent in the genomic regions
489 around each associated peaks.
490
491 We preferred this approach rather than using local LD extent in order to limit the effect of (i)
492 the strong variation of marker density along the chromosome (Additional file 2: Figure S2),
493 (ii) the local ascertainment bias due to the markers sampling (iii) the poor estimation of the
494 local recombination rate using a genetic map, notably for low recombination regions [3, 41]
495 (iv) errors in locus order due to assembly errors or chromosomal rearrangements.
496
497 We compared *LD_win* with *LD_adj*, another approach based on LD to group the SNPs
498 associated to trait variation into QTL. The discrepancies between the two approaches can be
499 explained by the local recombination rate and LD pattern. Since *LD_adj* approach was based
500 on the grouping of contiguous SNPs according to their LD, this approach was highly sensitive
501 to (i) error in marker order or position due to genome assembly errors or structural variations,
502 which are important in maize [44] (ii) genotyping or imputation errors, which we estimated at
503 *ca.* 1% and *ca.* 4%, respectively, for GBS (Additional file 1: Table S1), (iii) presence of
504 allelic series with contrasted effects in different experiments which are currently observed in
505 maize [40], (iv) LD threshold used. On the other hand, *LD_win* lead either to inflate the
506 number of QTLs in high recombinogenic regions in which SNPs were too distant genetically
507 to be grouped, or deflated their number by grouping associated SNPs in low recombinogenic

508 regions. Since *LD_win* considered the average LD extent, this method could conduct either to
509 separate or group abusively SNPs when local LD extent were different than the global LD
510 extent.
511
512 Note that LD windows should not be considered as confidence intervals since the
513 relationship between LD and recombination is complex due to demography, drift and
514 selection in association panels, contrary to linkage based QTL mapping [17]. The magnitude
515 of the effect of causal polymorphism in the estimation of these intervals which is well
516 established for linkage mapping should be explored further [45]. Other approaches have been
517 proposed to cluster SNPs according to LD [46, 47]. These approaches aim at segmenting the
518 genome in different haplotype blocks separating by high recombination regions. These
519 methods are difficult to use for estimating putative windows inside which the causal
520 polymorphisms are because such approaches are not centered on the associated SNP.
521 Several QTLs identified by *LD_win* in our study correspond to regions previously
522 identified: in particular six regions associated with female flowering time [27] and 30 regions
523 associated with different traits in the Cornfed dent panel [11]. Conversely, we did not identify
524 in our study any QTL associated to the florigen *ZCN8*, which showed significant effect in
525 these two previous studies. This relates most likely to the fact that we narrowed the flowering
526 time range in our study, in particular by eliminating early lines. This reduced the
527 representation of the early allele in the Zcn8, leading to a MAF of 0.27 in our study vs. 0.35 in
528 Rincent *et al.* [11], which can diminishes the power of the tests [14].

Complementarity of 600K and GBS for QTL detection resulted mostly from the tagging of different haplotypes rather than the coverage of different genomic regions.

Number of significant SNPs and QTLs increased with the increase in marker number (Table 3, Figure 4). This could be explained partly by a better coverage of some genomic regions by SNPs, notably in high recombinogenic regions which showed a very short LD extent and were enriched in QTLs (Additional file 10: Figure S13). Numerous new QTLs identified by the 600K array and GBS as compared with those identified by the 50K array were detected in high recombinogenic regions that were considerably less covered by the 50K array than the 600K array or GBS (Additional file 2: Figure S2).

The high complementarity for QTL detection between GBS and 600K array was only explained to a limited extent by the difference of the SNP distribution and density along the genome, since these two technologies targeted similar regions as showed by coverage analysis (Figure 2 and Additional file 2: Figure S2). However, at a finer scale, SNPs from the 600K array and GBS could tag close but different genomic regions around genes. SNPs from the 600K array were mostly selected within coding regions of genes [4], whereas SNP from GBS targeted more largely low copy regions, which included coding but also regulatory regions of genes [32, 35]. To further analyse the complementarity of the technologies, we analysed local haplotypes. We showed that both technologies captured different haplotypes when similar genomic regions were targeted (Figure 6). Hence, we pinpointed that GBS and DNA arrays are highly complementary for QTL detection because they tagged different haplotypes rather than tagging different regions (Figure 6). Based on the L-shaped MAF distribution, which suggest no ascertainment bias, and the high number of sequenced lines used for the GBS, we expect a closer representation of the variation present in our panel by this technology

compared to the 600K array, but this comes to the cost of an enrichment in rare alleles. Both factors tend to counterbalance each other in terms of GWAS power.

Our results suggest that we did not reach saturation with our *c.* 800,000 SNPs because (i) some haplotypes certainly remain not tagged (ii) the genome coverage was not complete, and (iii) the number of significant SNPs and QTLs continued to increase with marker density (Figure 4). Considering LD and marker density, the genotypic data presently available were most likely enough to well represent polymorphisms in the centromeric regions, whereas using more markers would be beneficial for telomeric regions. New approaches based on resequencing of representative lines and imputation are currently developed to achieve this goal.

Methods

Plant Material and Phenotypic Data

The panel of 247 genotypes (Additional file 11: Table S6) includes 164 lines from a wider panel of the CornFed project, composed of dent lines from Europe and America [11] and 83 additional lines derived from public breeding programs in Hungary, Italy and Spain and recent lines free of patent from the USA. Lines were selected within a restricted window of flowering time (10 days). Candidate lines with poor sample quality, i.e. high level of heterozygosity, or high relatedness with other lines were discarded. The lines selection was also guided by pedigree to avoid as far as possible over-representation of some parental materials. These inbred lines were crossed with a common flint tester (UH007) and the

572 hybrids were evaluated for male flowering time (Day To Anthesis, DTA), plant height
573 (plantHT), and grain yield (GY) at seven sites in Europe, during two years (2012 and 2013),
574 and for two water treatments (watered and rainfed) [30]. The adjusted mean (Best Linear
575 Unbiased Estimation, BLUEs, https://doi.org/10.15454/IASSTN) of the three traits were
576 estimated per environment (site × year × treatment) using a mixed model with correction for
577 blocks, repetitions and rows and columns in order to take into account spatial variation of
578 micro-environment in each field trial [30]. Variance components and heritability of each traits
579 in each environment were also estimated [30] (Additional file 12: Table S7). Adjusted means
580 of hybrids were combined with genotyping data of the lines to perform GWAS.

Genotyping and Genotyping-By-Sequencing Data

581 The inbred lines were genotyped using three technologies: a maize Illumina Infinium HD 50K
582 array [3], a maize Affymetrix Axiom 600K array [4], and Genotyping-By-Sequencing [2, 35].
583 In the arrays, DNA fragments are hybridized with probes attached to the array that flanked
584 SNPs that have been previously identified between inbred lines (Additional file 5:
585 Supplementary Text 1 for the description of the data from the two arrays). Genotyping-by-
586 sequencing technology is based on multiplex resequencing of tagged DNA from different
587 individuals for which some genomic regions were targeted using restriction enzyme (Keygene
588 N.V. owns patents and patent applications protecting its Sequence Based Genotyping
589 technologies) [2]. Cornell Institute (NY, USA) processed raw sequence data using a multi-
590 step Discovery and a one-top Production pipeline (*TASSEL-GBS*) in order to obtain genotypes
591 (Additional file 5: Supplementary Text 2). An imputation step of missing genotypes was
592 carried out by Cornell Institute [36], which utilized an algorithm that searches for the closest
593 neighbour in small SNP windows across the haplotype library [35], allowing for a 5%

mismatch. If the requirements were not met, the SNP was left ungenotyped for individuals.

We applied different filters (heterozygosity rate, missing data rate, minor allele frequency) for a quality control of the genetic data before performing the diversity and association genetic analyses. For GBS data, the filters were applied after imputation using the method "Compilation of Cornell homozygous genotypes and Beagle genotypes" (GBS_5 in Additional file 1: Figure S1; See section "Evaluating Genotyping and Imputation Quality"). We eliminated markers that had an average heterozygosity and missing data rate higher than 0.15 and 0.20, respectively, and a Minor Allele Frequency (MAF) lower than 0.01 for the diversity analyses and 0.05 for the GWAS. Individuals which had heterozygosity and/or missing data rate higher than 0.06 and 0.10, respectively, were eliminated.

Evaluating Genotyping and Imputation Quality

Estimating the genotyping and imputation quality were performed using 245 lines since two inbred lines have different seedlots between technologies. The 50K and the 600K arrays were taken as reference to compare the concordance of genotyping (genotype matches) with the imputation of GBS based on their position. While SNP position and orientation from GBS were called on the reference maize genome B73 AGP_v2 (release 5a) [48], flanking sequences of SNPs in the 50K array were primary aligned on the first maize genome reference assembly B73 AGP_v1 (release 4a.53) [49]. Both position and orientation scaffold carrying SNPs from the 50K array can be different in the AGP_v2, which could impair correct comparison of genotype between the 50K array and GBS. Hence, we aligned flanking sequences of SNPs from the 50K array on maize B73 AGP_v2 using the Basic Local

Alignment Search Tool (BLAST) to retrieve both positions and genotype in the same and correct strand orientation (forward) to compare genotyping. The number of common markers between the 50K/600K, 50K/GBS, GBS/600K and 50K/600K/GBS was 36,395, 7,018, 25,572 and 5,947 SNPs, respectively. The comparison of the genotyping and imputation quality between the 50K/GBS, 50K/600K and 600K/GBS was done on 5,336 and 24,286 PANZEA markers [50] in common, and 26,154 markers in common, respectively. The genotyping concordance of the 600K with the 50K array was extremely high (99.50%) but slightly lower for heterozygotes (92.88%). In order to achieve these comparisons, we considered the direct reads from GBS (**GBS_1**) and four approaches for imputation (GBS_2 to GBS_5, Additional file 1: Figure S1). **GBS_2** approach consisted in one imputation step from the direct read by Cornell University, using *TASSEL* software, but missing data was still present. **GBS_3** approach consisted in a genotype imputation of the whole missing data of the direct read by *Beagle v3* [13]. In **GBS_4,** genotype imputation by Beagle was performed on Cornell imputed data after replacing the heterozygous genotypes into missing data. **GBS_5**, consisted in homozygous genotypes of GBS_2 completed by values imputed in GBS_3 (Additional file 1: Figure S1).

Diversity Analyses

After excluding the unplaced SNPs and applying the filtering criteria for the diversity analyses (MAF > 0.01), we obtained the final genotyping data of the 247 lines with 44,729 SNPs from the 50K array, 506,662 SNPs from the 600K array, and 395,024 SNPs from the GBS (Figure 1). All markers of the 600K array and GBS_5 that passed the quality control were used to perform the diversity analyses (estimation of Q genetic groups and K kinships). For the 50K, we used only the PANZEA markers (29,257 SNPs) [50] in order to reduce the

638 ascertainment bias noted by Ganal *et al.* [3] when estimating Nei's index of diversity [51] and
639 relationship coefficients. Genotypic data generated by the three technologies were organized
640 as *G* matrices with *N* rows and *L* columns, *N* and *L* being the panel size and number of
641 markers, respectively. Genotype of individual *i* at marker *l* ($G_{i,l}$) was coded as 0 (the
642 homozygote for an arbitrarily chosen allele), 0.5 (heterozygote), or 1 (the other homozygote).
643 Identity-By-Descend (IBD) was estimated according to Astle and Balding [19]:

644 $$K\_Freq_{i,j} = \frac{1}{L}\sum_{l=1}^{L}\frac{(G_{i,l}-p_l)(G_{j,l}-p_l)}{p_l(1-p_l)},$$

645 where p_l is the frequency of the allele coded 1 of marker *l* in the panel of interest, *i*
646 and *j* indicate the inbred lines for which the kinship was estimated. We also estimated the
647 Identity-By-State (IBS) by estimating the proportion of shared alleles. For GWAS, we used
648 *K_Chr* [14] that are computed using similar formula as *K_Freq*, but with the genotyping data
649 of all the chromosomes except the chromosome of the SNP tested. This formula provides an
650 unbiased estimate of the kinship coefficient and weights by allelic frequency assuming Hardy-
651 Weinberg equilibrium. Hence, relatedness is higher if two individuals share rare alleles than
652 common alleles.

653

654 Genetic structure was analysed using the sofware *ADMIXTURE v1.22* [18] with a
655 number of groups varying from 2 to 10 for the three technologies. We compared assignation
656 by *ADMIXTURE* of inbred lines between the three technologies by estimating the proportion
657 of inbred lines consistently assigned between technologies two by two (50K *vs* GBS_5, 50K *vs*
658 600K, 600K *vs* GBS_5) using a threshold of 0.5 for admixture.

659

660 Expected heterozygosity (He) [51] was estimated at each marker as $2p_l(1 - p_l)$ and
661 was averaged on all the markers for a global characterization of the panel for the three

662 technologies. Principal Coordinate Analyses (PCoA) were performed on the genetic distance
663 matrices [52], estimated as $1_{N,N} - K\_Freq$, where $1_{N,N}$ is a matrix of ones of the same size as
664 *K_Freq*.

Linkage Disequilibrium Analyses

665 We first analyzed the effect of the genetic structure and kinship on linkage disequilibrium
666 (LD) extent within and between chromosomes by estimating genome-wide linkage
667 disequilibrium using the 29,257 PANZEA SNPs from the 50K array. Four estimates of LD
668 were used: the squared correlation (r^2) between allelic dose at two markers [53], the squared
669 correlation taking into account global kinship with *K_Freq* estimator (r^2K), the squared
670 correlation taking into account population structure (r^2S), and the squared correlation taking
671 into account both (r^2KS) [17].
672
673 To explore the variation of LD decay and the stability of LD extent along the chromosomes,
674 we estimated LD between a non-redundant set of 810,580 loci from the GBS, the 50K and
675 600K arrays. To save computation time, we calculated LD between loci within a sliding
676 window of 1 cM. Genetic position was obtained by projecting the physical position of each
677 locus using a *smooth.spline* function R calibrated on the genetic consensus map of the
678 Cornfed Dent Nested Association Mapping (NAM) design [40]. We used the estimator r^2 and
679 r^2K using 10 different kinships *K_Chr*. This last estimator was calculated because it
680 corresponds exactly to LD used to map QTL in our GWAS model. It determines the power of
681 GWAS to detect QTL considering that causal polymorphisms were in LD with some
682 polymorphisms genotyped in our panel [17]. To study LD extent variation, we estimated LD
683 extent by adjusting Hill and Weir's model [37] using non-linear regression (*nls* function in R-

684 package *nlme*) against both physical and genetic position within each chromosome. Since
685 recombination rate (cM / Mbp) varied strongly along the genome (Figure 2 and Additional
686 file 2: Figure S2), we defined high (>0.5 cM / Mbp) and low (<0.5 cM / Mbp)
687 recombinogenic genomic regions within each chromosome. We adjusted Hill and Weir's
688 model [37] separately in low and high recombinogenic regions (Additional file 10: Table S5)
689 by randomly sampling 100 sets of 500,000 pairs of loci distant from less than 1 cM. This
690 random sampling avoided over-representation of pairs of loci from low recombinogenic
691 regions due to the sliding-window approach (Additional file 12: Figure S14). 500,000 pairs of
692 loci represented 0.36% (Chromosome 3 / High rec) to 1.20% of all pairs of loci (Chromosome
693 8 / High rec).
694 For all analyses, we estimated LD extent by calculating the genetic and physical distance for
695 the fitted curve of Hill and Weir's Model that reached r^2K=0.1, r^2K=0.2 and r^2K=0.4.

Genome coverage estimation

696 In order to estimate the genomic regions in which the effect of an underlying causal
697 polymorphisms could be captured by GWAS using LD with SNP from three technologies, we
698 developed an approach to define LD windows around each SNP with MAF ≥ 5% based on
699 LD extent (Additional file 12: Figure S14). To set the LD window around each SNP, we used
700 LD extent with r^2K=0.1 (negligible LD), r^2K=0.2 (intermediate LD) and r^2K=0.4 (high LD)
701 estimated in low and high recombinogenic regions for each chromosome. We used the global
702 LD decay estimated for these large chromosomal regions rather than local LD extent (i) to
703 avoid bias due to SNP sampling within small genomic regions, (ii) to reduce computational
704 time, and (iii) to limit the impact of possible local error in genome assembly. In low
705 recombinogenic regions, we used the physical LD extent, hypothesizing that recombination

rate is constant along physical distance in these regions. In high recombinogenic regions, we used the genetic LD extent since there is a strong variation of recombination rate by base pair along the physical position (Additional file 2: Figure S2). We then converted genetic LD windows into physical windows by projecting the genetic positions on the physical map using the *smooth.spline* function implemented in R calibrated on the NAM dent consensus map [40]. Reciprocally, we obtained the genetic positions of LD windows in low recombinogenic regions by projecting the physical boundaries of LD windows on the genetic map.

To estimate coverage of the three technologies to detect QTLs based on their SNP distribution and density, we calculated cumulative genetic and physical length that are covered by LD windows around the markers, considering different LD extents for each chromosome (r^2K=0.1, r^2K=0.2, r^2K=0.4). In order to explore variation of genome coverage along the chromosome, we estimated the proportion of genome covered using a sliding-windows approach based on physical distance (2Mbp).

Statistical Models for Association Mapping

We used four models to determine the statistical models that control best the confounding factors (*i.e.* population structure and relatedness) in GWAS (Additional file 5: Supplementary Texts 3 and 4). We tested different software implementing either approximate (EMMAX) [8] or exact computation of standard test statistics (ASReml and FaST-LMM) [6, 54] for computational time and GWAS results differences (Additional file 5: Supplementary Text 5). Single-trait, single-environment GWAS was performed for each marker for each environment and all traits using FaST-LMM. We selected the mixed model using *K_Chr*, estimated from PANZEA markers of the 50K array to perform GWAS on 66 situations (environment × trait)

(Additional file 5: Supplementary Text 4, Additional file 12: Figure S15 and Additional file 12: Figure S16). We developed a GWAS pipeline in *R v3.2.1* [55] calling FaST-LMM software and implementing [14] approaches to conduct single trait and single environment association tests.

Multiple testing is a major challenge in GWAS using large numbers of markers. The experiment-wise error rate (α_e) increases with the number of tests (number of markers) carried out, even when the point-wise error rate (α_p) is maintained low. Popular methods [56, 57] are overly conservative and can result in overlooking true positive associations. In addition, these corrections assume that the hypothesis tests are independent. To take into account the dependence of the tests in GWAS, α_p has to be adjusted in order to keep α_e at a nominal level. Moskvina and Schmidt [58] and Gao *et al.* [59, 60] corrections can correctly infer the number of independent tests and use the Bonferroni formula to rapidly adjust for multiple testing. Using Gao approaches, we estimated the number of independent tests for GWAS at 15,780 for the 50K, 92,752 for the 600K, 109,117 for the GBS_5 and 191,026 for the combined genetic data, leading to different $-log_{10}(p\text{-}value)$ thresholds: 5.49, 6.27, 6.34 and 6.58, respectively. Because of these differences, we used two thresholds of $-log_{10}(p\text{-}value) = 5$ (less stringent) *and 8* (hightly conservative and slightly above Bonferroni) for comparing GWAS to avoid the differences of identification of significant SNPs between the technologies due to the choice of the threshold.

Methods for grouping associated SNPs into QTLs

We used two approaches based on LD for grouping significant SNPs. The first approach (*LD_win*) used LD windows, previously described, to group significant SNPs into QTLs considering that all SNPs with overlapping LD windows of $r^2K=0.1$ belong to the same QTL. We hypothesized that significant SNPs with overlapping LD windows at $r^2K=0.1$ captured the same causal polymorphism and were therefore a single and unique QTL. The second approach (*LD_adj*) grouped into single QTL significant SNPs that are adjacent on the physical map providing that their LD were above a LD threshold ($r^2K > 0.5$). We used LD heatmaps for comparing the SNP grouping produced by the two approaches on the three different traits across all environments (Additional file 6: Figure S9-LD-Adjacent and Additional file 7: Figure S9-LD-Windows). All scripts are implemented in R software [55].

List of abreviations

DTA = Day to Anthesis

GY = Grain Yield adjusted at 15% moisture

plantHT = Plant Height

GBS = Genotyping By Sequencing

LD = Linkage disequilibrium

GWAS = Genome-Wide Association Studies

MAF = Minimum Allelic Frequency

SNP = Single Nucleotide Polymorphism

HRR = High Recombinogenic Regions

LRR = Low Recombinogenic Regions

QTL = Quantitative Trait Locus

771

772 Declarations

Ethics approval and consent to participate

773 Not applicable.

Consent for publication

774 Not applicable.

Availability of data and material

775 The following links toward the data will be available upon publication of this paper.

776 All the genotyping data used in this study can be found at https://doi.org/10.15454/AEC4BN.

777 The GWAS results can be found at https://doi.org/10.15454/6TL2N4.

778 The phenotypic dataset can be found at https://doi.org/10.15454/IASSTN.

Competing interests

779 The authors declare that they have no competing interests.

Funding

780 This project (Project ID: 244374) was funded under the European FP7- KBBE (CP – IP –
781 Large-scale integrating project, DROPS) and the *Agence Nationale de la Recherche* project
782 ANR-10-BTBR-01 (ANR-PIA AMAIZING).

Authors' contributions

S.S.N., S.D.N. and A.C., designed the studied and wrote the article. S.S.N. performed genotyping data quality control, imputation and genetic analyses. S.D.N. developed and performed LD analyses. A.C. designed the association panel with the help of S.D.N. and C.W. C.B. participated in assembling the dent inbred lines panel, organizing the germplasms and field work for seeds production. E.J.M., C.W. and F.T. collected and analysed the phenotypic data. V.C. and D.M. performed DNA extraction and prepared the samples. All authors critically reviewed and approved the final manuscript.

Acknowledgements

We are grateful to key partners from the field: Pierre Dubreuil, Cécile Richard, Jérémy Lopez (Biogemma), Tamás Spitkó (MTA ATK), Therese Welz (KWS), Franco Tanzi, Ferenc Racz, Vincent Schlegel (Syngenta) and Maria Angela Canè (UNIBO). We also acknowledge Björn Usadel and Axel Nagel (MPI) for data management. We thank Willem Kruijer, Fred Van Eeuwijk (WUR), Tristan Mary-Huard and Laurence Moreau (INRA) for helpful discussions and statistical advice. We are grateful to Chris-Carolin Schön (TUM) for providing an early access to the Affymetrix Axiom 600K array and Edward Buckler (USDA) for providing genotyping using GBS. We are also grateful to partners of the CornFed project, Univ. Hohenheim (Germany), CSIC (Spain), CRAG (Spain), MTA ATK (Hungary), NCRPIS (USA), CRB Maize (France) and CRA-MAC (Italy) who contributed to the genetic material.

Authors' information (optional)

Not applicable.

References

1. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES: **Genome-wide association study of leaf architecture in the maize nested association mapping population.** *Nature Genetics* 2011, **43:**159.

2. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE: **A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species.** *PLoS ONE* 2011, **6:** e19379.

3. Ganal MW, Durstewitz G, Polley A, Bérard A, Buckler ES, Charcosset A, Clarke JD, Graner E-M, Hansen M, Joets J, et al: **A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome.** *PLoS ONE* 2011, **6:** e28334.

4. Unterseer S, Bauer E, Haberer G, Seidel M, Knaak C, Ouzunova M, Meitinger T, Strom TM, Fries R, Pausch H, et al: **A powerful tool for genome analysis in maize: development and evaluation of the high density 600 k SNP genotyping array.** *BMC Genomics* 2014, **15:**823.

5. Zhou X, Stephens M: **Genome-wide efficient mixed-model analysis for association studies.** *Nature Genetics* 2012, **44:**821.

6. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D: **FaST linear mixed models for genome-wide association studies.** *Nature Methods* 2011, **8:**833.

7. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D: **Improved**

linear mixed models for genome-wide association studies. *Nature Methods* 2012, **9:**525.

8. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-y, Freimer NB, Sabatti C, Eskin E: **Variance component model to account for sample structure in genome-wide association studies.** *Nature Genetics* 2010, **42:**348.

9. Flint-Garcia SA, Thornsberry JM, Buckler ES: **Structure of Linkage Disequilibrium in Plants.** *Annual Review of Plant Biology* 2003, **54:**357-374.

10. Remington DL, Thornsberry JM, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES: **Structure of linkage disequilibrium and phenotypic associations in the maize genome.** *Proceedings of the National Academy of Sciences* 2001, **98:**11479.

11. Rincent R, Nicolas S, Bouchet S, Altmann T, Brunel D, Revilla P, Malvar RA, Moreno-Gonzalez J, Campo L, Melchinger AE, et al: **Dent and Flint maize diversity panels reveal important genetic potential for increasing biomass production.** *Theoretical and Applied Genetics* 2014, **127:**2313-2331.

12. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, et al: **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness.** *Nature Genetics* 2006, **38:**203.

13. Browning SR, Browning BL. **Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering.** The American Journal of Human Genetics 2007, **81** (5): 1084-1097

14. Rincent R, Moreau L, Monod H, Kuhn E, Melchinger AE, Malvar RA, Moreno-

Gonzalez J, Nicolas S, Madur D, Combes V, et al: **Recovering Power in Association Mapping Panels with Variable Levels of Linkage Disequilibrium.** *Genetics* 2014, **197:**375.

15. Van Inghelandt D, Melchinger AE, Lebreton C, Stich B: **Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers.** *Theoretical and Applied Genetics* 2010, **120:**1289-1299.

16. Nicolas SD, Péros J-P, Lacombe T, Launay A, Le Paslier M-C, Bérard A, Mangin B, Valière S, Martins F, Le Cunff L, et al: **Genetic diversity, linkage disequilibrium and power of a large grapevine (Vitis vinifera L) diversity panel newly designed for association studies.** *BMC Plant Biology* 2016, **16:**74.

17. Mangin B, Siberchicot A, Nicolas S, Doligez A, This P, Cierco-Ayrolles C: **Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness.** *Heredity* 2012, **108:**285.

18. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals.** *Genome Research* 2009, **19:**1655-1664.

19. Astle W, Balding DJ: **Population Structure and Cryptic Relatedness in Genetic Association Studies.** *Statist Sci* 2009, **24:**451-471.

20. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: **Association Mapping in Structured Populations.** *The American Journal of Human Genetics* 2000, **67:**170-181.

21. VanRaden PM: **Efficient Methods to Compute Genomic Predictions.** *Journal of Dairy Science* 2008, **91:**4414-4423.

22. Bernardo R: **Genomewide Markers for Controlling Background Variation in Association Mapping.** *The Plant Genome* 2013, **6**.

873 23. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES:
874 **Dwarf8 polymorphisms associate with variation in flowering time.** *Nature*
875 *Genetics* 2001, **28:**286.
876 24. Hufford MB, Lubinksy P, Pyhäjärvi T, Devengenzo MT, Ellstrand NC, Ross-Ibarra J:
877 **The Genomic Signature of Crop-Wild Introgression in Maize.** *PLOS Genetics*
878 2013, **9:**e1003477.
879 25. Rincent R, Laloë D, Nicolas S, Altmann T, Brunel D, Revilla P, Rodríguez VM,
880 Moreno-Gonzalez J, Melchinger A, Bauer E, et al: **Maximizing the Reliability of**
881 **Genomic Selection by Optimizing the Calibration Set of Reference Individuals:**
882 **Comparison of Methods in Two Diverse Groups of Maize Inbreds (*Zea Mays L.*).**
883 *Genetics* 2012, **192:**715.
884 26. Bouchet S, Bertin P, Presterl T, Jamin P, Coubriche D, Gouesnard B, Laborde J,
885 Charcosset A: **Association mapping for phenology and plant architecture in maize**
886 **shows higher power for developmental traits compared with growth influenced**
887 **traits.** *Heredity* 2017, **118:**249-259.
888 27. Bouchet S, Servin B, Bertin P, Madur D, Combes V, Dumas F, Brunel D, Laborde J,
889 Charcosset A, Nicolas S: **Adaptation of Maize to Temperate Climates: Mid-**
890 **Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key**
891 **Genomic Regions, with a Major Contribution of the Vgt2 (ZCN8) Locus.** *PLoS*
892 *ONE* 2013, **8:**e71377.
893 28. Messing J, Dooner HK: **Organization and variability of the maize genome.** *Current*
894 *Opinion in Plant Biology* 2006, **9:**157-163.
895 29. Hu H, Schrag TA, Peis R, Unterseer S, Schipprack W, Chen S, Lai J, Yan J, Prasanna
896 BM, Nair SK, et al: **The Genetic Basis of Haploid Induction in Maize Identified**

with a Novel Genome-Wide Association Method. *Genetics* 2016, **202:**1267.

30. Millet EJ, Welcker C, Kruijer W, Negro S, Coupel-Ledru A, Nicolas SD, Laborde J, Bauland C, Praud S, Ranc N, et al: **Genome-Wide Analysis of Yield in Europe: Allelic Effects Vary with Drought and Heat Scenarios.** *Plant Physiology* 2016, **172:**749.

31. Crossa J, Beyene Y, Kassa S, Pérez P, Hickey JM, Chen C, de los Campos G, Burgueño J, Windhausen VS, Buckler E, et al: **Genomic Prediction in Maize Breeding Populations with Genotyping-by-Sequencing.** *G3: Genes|Genomes|Genetics* 2013, **3:**1903.

32. Romay MC, Millard MJ, Glaubitz JC, Peiffer JA, Swarts KL, Casstevens TM, Elshire RJ, Acharya CB, Mitchell SE, Flint-Garcia SA, et al: **Comprehensive genotyping of the USA national maize inbred seed bank.** *Genome Biology* 2013, **14:**R55.

33. Gouesnard B, Negro S, Laffray A, Glaubitz J, Melchinger A, Revilla P, Moreno-Gonzalez J, Madur D, Combes V, Tollon-Cordet C, et al: **Genotyping-by-sequencing highlights original diversity patterns within a European collection of 1191 maize flint lines, as compared to the maize USDA genebank.** *Theoretical and Applied Genetics* 2017, **130:**2165-2189.

34. Frascaroli E, Schrag TA, Melchinger AE: **Genetic diversity analysis of elite European maize (Zea mays L.) inbred lines using AFLP, SSR, and SNP markers reveals ascertainment bias for a subset of SNPs.** *Theoretical and Applied Genetics* 2013, **126:**133-141.

35. Glaubitz JC, Casstevens TM, Lu F, Harriman J, Elshire RJ, Sun Q, Buckler ES: **TASSEL-GBS: A High Capacity Genotyping by Sequencing Analysis Pipeline.** *PLoS ONE* 2014, **9:**e90346.

36. Swarts K, Li H, Romero Navarro JA, An D, Romay MC, Hearne S, Acharya C, Glaubitz JC, Mitchell S, Elshire RJ, et al: **Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants.** *The Plant Genome* 2014, **7**.

37. Hill WG, Weir BS: **Variances and covariances of squared linkage disequilibria in finite populations.** *Theoretical Population Biology* 1988, **33:**54-78.

38. Charlesworth D, Willis JH: **The genetics of inbreeding depression.** *Nature Reviews Genetics* 2009, **10:**783.

39. Hudson RR, Kaplan NL: **Deleterious background selection with recombination.** *Genetics* 1995, **141:**1605.

40. Giraud H, Bauland C, Falque M, Madur D, Combes V, Jamin P, Monteil C, Laborde J, Palaffre C, Gaillard A, et al: **Reciprocal Genetics: Identifying QTL for General and Specific Combining Abilities in Hybrids Between Multiparental Populations from Two Maize (*Zea mays L.*) Heterotic Groups.** *Genetics* 2017, **207:**1167.

41. Bauer E, Falque M, Walter H, Bauland C, Camisan C, Campo L, Meyer N, Ranc N, Rincent R, Schipprack W, et al: **Intraspecific variation of recombination rate in maize.** *Genome Biology* 2013, **14:**R103.

41. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES: **Genome-wide association study of leaf architecture in the maize nested association mapping population.** *Nature Genetics*

2011, **43:**159.

42. Le Gouis J, Bordes J, Ravel C, Heumez E, Faure S, Praud S, Galic N, Remoué C, Balfourier F, Allard V, Rousset M: **Genome-wide association analysis to identify chromosomal regions determining components of earliness in wheat.** *Theoretical and Applied Genetics* 2012, **124:**597-611.

43. Cormier F, Le Gouis J, Dubreuil P, Lafarge S, Praud S: **A genome-wide identification of chromosomal regions determining nitrogen use efficiency components in wheat (Triticum aestivum L.).** *Theoretical and Applied Genetics* 2014, **127:**2679-2693.

44. Springer NM, Ying K, Fu Y, Ji T, Yeh C-T, Jia Y, Wu W, Richmond T, Kitzman J, Rosenbaum H, et al: **Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content.** *PLOS Genetics* 2009, **5:**e1000734.

45. Darvasi A, Weinreb A, Minke V, Weller JI, Soller M: **Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map.** *Genetics* 1993, **134:**943.

46. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al: **The Structure of Haplotype Blocks in the Human Genome.** *Science* 2002, **296:**2225.

47. Wang H, Chu WS, Hemphill C, Elbein SC: **Human Resistin Gene: Molecular Scanning and Evaluation of Association with Insulin Sensitivity and Type 2 Diabetes in Caucasians.** *The Journal of Clinical Endocrinology & Metabolism* 2002, **87:**2520-2524.

48. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J,

Fulton L, Graves TA, et al: **The B73 Maize Genome: Complexity, Diversity, and Dynamics.** *Science* 2009, **326:**1112.

49. Ganal MW, Altmann T, Röder MS: **SNP identification in crop plants.** *Current Opinion in Plant Biology* 2009, **12:**211-217.

50. Gore MA, Chia J-M, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL, Peiffer JA, McMullen MD, Grills GS, Ross-Ibarra J, et al: **A First-Generation Haplotype Map of Maize.** *Science* 2009, **326:**1115.

51. Nei M: **Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals.** *Genetics* 1978, **89:**583.

52. Gower JC: **Some distance properties of latent root and vector methods used in multivariate analysis.** *Biometrika* 1966, **53:**325-338.

53. Hill WG, Robertson A: **Linkage disequilibrium in finite populations.** *Theoretical and Applied Genetics* 1968, **38:**226-231.

54. Butler DG, Cullis BR, Gilmour AR, Gogel BJ: **ASReml-R reference manual.** *The State of Queensland, Department of Primary Industries and Fisheries, Brisbane, Australia* 2009.

55. R Core Team: **R: A language and environment for statistical computing.** *R Foundation for Statistical Computing, Vienna, Austria* 2015**:**URL https://www.R-project.org/.

56. Bonferroni CE: **Teoria statistica delle classi e calcolo delle probabilità.** *Pubblicazioni dell'Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 1936, **8:**3-62.

57. Šidák Z: **Rectangular Confidence Regions for the Means of Multivariate Normal Distributions.** *Journal of the American Statistical Association* 1967, **62:**626-633.

58. Moskvina V, Schmidt KM: **On multiple-testing correction in genome-wide association studies.** *Genetic Epidemiology* 2008, **32:**567-573.

59. Gao X, Becker LC, Becker DM, Starmer JD, Province MA: **Avoiding the high Bonferroni penalty in genome-wide association studies.** *Genetic Epidemiology* 2010, **34:**100-105.

60. Gao X, Starmer J, Martin ER: **A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms.** *Genetic Epidemiology* 2008, **32:**361-369.

Figure legends

Figure 1: Comparison of genotyping data between 50K and 600K arrays, and GBS. (a**)** Distribution of minor allele frequency per SNP before filtering (monomorphic SNPs removed). (b**)** Distribution of SNP missing data proportion for the 50K array, 600K array, GBS direct reads (GBS_1) and GBS after imputation by Cornell Institute (GBS_2, note that the scale of the x-axes is different). (**c**) Relatedness distribution (Identity-By-State, IBS) after QC filtering with MAF≥1% (IBS using GBS_1 was not estimated because of the low calling rate).

Figure 2: Variation of the markers density (top), the recombination rate (middle) and the genome coverage (bottom). Non-overlapping 2Mbp windows along the chromosome 3 were used. The percentage of genome coverage used the cumulative length of LD windows calculated around each SNP. Markers along chromosome 3 have MAF ≥ 5%. Green, blue, red and black lines represent variation of GBS, 600K, 50K and combined technologies,

respectively.

Figure 3: Correlation (*r*) of the Identity-By-State (IBS) between the three technologies (after imputation). (**a**) IBS_{600K} vs IBS_{50K}, (**b**) IBS_{GBS} vs IBS_{50K}, (**c**) IBS_{GBS} vs IBS_{600K}. The red line indicates the bisector.

Figure 4: Number of significant SNPs (blue line) and QTLs (red line) identified as a function of SNP density (x-axis) for the three traits (DTA, male flowering time; plantHT, plant height; GY, grain yield).

Figure 5: Complementarity of the three technologies to detect QTLs. The numbers of specific QTLs detected by each technologies for the three traits (flowering time, plant height, grain yield) are shown.

Figure 6: Complementarity of QTLs detection between the 600K array and the GBS for two regions (QTL 32/QTL95). (**a**) Manhattan plot of the $-log_{10}$(*p*-value) along the genome. Dotted red lines correspond to QTL32 and QTL95 located on chromosome 1 and 3, respectively, for the flowering time in one environment (Ner13R). (**b**) Local manhattan plot of the $-log_{10}$(*p-value*) (top) and linkage disequilibrium corrected by the kinship (r^2K) (bottom) of all SNPs with the strongest associated marker within QTL 32 (left) and QTL 95 (right). (**c**) Local haplotypes displayed by SNPs within the QTLs 32 (left) and 95 (right). Inbred lines are in rows and SNPs are in columns. Inbred lines were ordered by hierarchical clustering based on local dissimilarity estimated by all SNPs within each QTLs. Genotyping matrix is colored according to their allelic dose at each SNP. Red and black represent homozygotes and gray represent heterozygotes. The associated peaks (red vertical lines) and other associated SNPs with $-log_{10}(p\text{-}value) > 5$ (orange vertical lines) are indicated above the genotyping matrix. H1,

1040 H2, H3, H4, H5 represent the 5 and 3 haplotypes obtained by cutting the dendograms with the
1041 most 5 and 3 dissimilar clusters within QTL32 and QTL95, respectively.
1042

1043 ## Additional file legends

1044 **Additional file 1 (.docx):**

1045 **Figure S1:** Different approaches used to compare the quality of genotyping and imputation of
1046 the GBS. We considered the direct reads from GBS (**GBS_1**) and four approaches for
1047 imputation (GBS_2 to GBS_5). **GBS_2** approach consisted in one imputation step from the direct
1048 read by Cornell University, using *TASSEL* software, but missing data was still present. **GBS_3**
1049 approach consisted in a genotype imputation of the whole missing data of the direct read by
1050 *Beagle v3*. In **GBS_4,** genotype imputation by Beagle was performed on Cornell imputed data
1051 after replacing the heterozygous genotypes into missing data. **GBS_5**, consisted in homozygous
1052 genotypes of GBS_2 completed by values imputed in GBS_3.
1053 **Table S1:** Percentage of GBS concordance based on the 50K and 600K arrays (Reference).
1054 Call rate of SNPs from GBS are in brackets. * After Beagle inference of missing data, the call
1055 rate is 100%. Here the call rate is <100% because the comparison was made against the 50K
1056 and the 600K arrays that include few missing data.
1057

1058 **Additional file 2 (.pdf):**

1059 **Figure S2:** Variation of the markers density, the recombination rate and the genome coverage
1060 in non-overlapping 2 Mbp windows along each chromosome. The percentage SNP coverage
1061 (bottom) used the cumulated length of physical LD windows around each SNP. Markers have
1062 MAF ≥ 5%. Green, blue, red and black lines represent variation of GBS, 600K, 50K and

combined technologies, respectively.

Additional file 3 (.docx):

Figure S3: Contribution of four ancestral populations to 247 inbred lines after ADMIXTURE analysis. Markers from the 50K array (top), 600K array (middle) and GBS (bottom) were used. One vertical bar corresponds to one individual. Lines were ordered according to contributions observed for the 50K array. From left to right, we have Stiff Stalk lines type B73 and B14a (red), Iodent lines type PH207 (green), Lancaster lines type Mo17 and Oh43 (turquoise), a group of lines assembling W117, F7057 type lines (blue).

Table S2: Means and ranges of the two relatedness estimators (IBS and IBD *i.e. K_Freq*) from the 50K (29,257 PANZEA SNPs only) and 600K arrays, and GBS.

Figure S4: Correlation (*r*) between the IBS and IBD (*K_Freq*) for each technology (A) and correlation of IBD between the three technologies (B). (C) Correlation of IBD between the three technologies after removing the excess of rare alleles in the GBS to have the same distribution of MAF as in the 50K and the 600K arrays. The red line is the bisector.

Figure S5: Principal coordinate analyses (PCoA) of the DROPS panel. The PCoA were based on the covariance matrix *K_Freq* estimated from the 50K Illumina array. The genetic groups identified by ADMIXTURE (N_Q = 4) are colored (differently than in Fig. S6). Three key founders are indicated (Iodent: PH207 in red, Stiff Stalk: B73 in blueviolet, Lancaster: Mo17 in turquoise).

Additional file 4 (.docx):

Figure S6: Heatmap of genome-wide linkage disequilibrium (LD) between all markers within and between chromosomes using PANZEA SNPs from the 50K array. All SNPs were ordered

1087 according to their position on the genome. Dots represented LD between two loci and were
1088 colored according to their strength. Classical LD measurement r^2 between loci were
1089 represented within triangle below the diagonal. Linkage disequilibrium corrected for structure
1090 (r^2S, A), relatedness (r^2K, B) or both (r^2KS, C) were represented within triangle above the
1091 diagonal.

1092 **Figure S7:** Linkage disequilibrium (r^2, top) and LD corrected for relatedness (r^2k, bottom) as
1093 a function of physical distance (left) and genetic distance (right): example of chromosome 1.

1094 **Figure S8:** Variation of genetic LD extent (Dm, cM), effective population size (N), along the
1095 physical map. A sliding window of 1 cM moving by 0.5 cM at each step was used. Local
1096 genetic LD extent (cM) and local effective size (N) were estimated by adjusting the Hill and
1097 Weir model's using r^2K between all loci that are located in sliding windows of 1 cM. Each
1098 values were plotted on the physical map of each chromosome by projecting the genetic
1099 position of the windows on the physical map.

1100

1101 **Additional file 5 (.docx):**

1102 **Supplementary Text 1:** Differences between microarrays.

1103 **Supplementary Text 2:** GBS pipelines.

1104 **Supplementary Text 3:** Statistical models for GWAS.

1105 **Supplementary Text 4:** Effects of confounding factors on GWAS.

1106 **Supplementary Text 5:** Performance of different software.

1107

1108 **Additional file 6 (.pdf):**

1109 **Figure S9-LD_Windows:** QTL limits obtained by the *LD_win* approach projected on
1110 heatmaps representing the level of LD between associated SNPs for each trait (DTA: male

flowering time, plantHT: plant height and GY: grain yield) and each chromosome. Upper and lower triangles on the heatmaps represented the r^2 and r^2K values between associated SNPs, respectively. Linkage disequilibrium between loci was colored according to values from weak LD (yellow) to high LD (red). The significant markers were ordered according to their physical positions on the chromosome and were represented by ticks on the four sides of the heatmaps. Limits of QTLs were displayed by gray dotted lines. QTL numbers were indicated in gray on the top and the right of each heatmap.

Additional file 7 (.pdf):

Figure S9-LD_Adjacent: QTL limits obtained by the *LD_Adj* approach projected on heatmaps representing the level of LD between associated SNPs for each trait (DTA: male flowering time, plantHT: plant height and GY: grain yield) for each chromosome. Upper and lower triangles on the heatmaps represented the r^2 and r^2K values between associated SNPs, respectively. Linkage disequilibrium between loci was colored according to values from weak LD (yellow) to high LD (red). The significant markers were ordered according to their physical positions on the chromosome and were represented by ticks on the four sides of the heatmaps. Limits of QTLs were displayed by gray dotted lines. QTL numbers were indicated in gray on the top and the right of each heatmap.

Additional file 8 (.pdf):

Table S3: Summary of all the QTLs identified for the male flowering time (DTA), plant height (plantHT) and grain yield (GY). “LowerLimit” and “UpperLimit” columns are the lower and upper physical limits for each QTL. The “Rec” column indicates if the QTL is located in a high or low region of recombination. “NbSNP50”, “LogPvaMax50”,

"NbSNP600", "LogPvaMax600", "NbSNPGBS", "LogPvaMaxGBS" are the number of significant SNPs and the most significant $-log_{10}(Pval)$ within the QTL for each technology across all environments. The physical position ("PosMax"), the proportion of the variance explained ("R2_LDMax") and the effect ("EffectMax") of the most significant SNP within the QTL is shown. "NbDiffEnv" gives the number of different situations that detected the QTL.

Additional file 9 (.docx):

Figure S10: Examples of comparison of QTLs detection on Chromosome 1, 6 and 8 for the different traits. Local distribution of the $-log_{10}(p\text{-}value)$ and linkage disequilibrium (bottom) corrected by the kinship (r^2k) of all SNPs with the strongest associated marker within the chosen QTL for the three technologies. Ticks on different x-axes show the marker density of the three technologies (red for the 50K, blue for the 600K and green for the GBS). The vertical red line spots the position of the SNP with the maximum $-log_{10}(p\text{-}value)$ within the QTL.

Additional file 10 (.docx):

Figure S11: Pleiotropy of QTLs between the traits. Number of QTLs specific and shared by the three traits across all environments. Note that several QTLs from one trait were sometimes included in a single QTL of another trait.

Figure S12: Percentage of stable QTLs across environments for the three traits (DTA: male flowering time, plantHT: Plant Height, GY: Grain Yield).

Table S4: Stability of QTL across environments. DTA: male flowering time, plantHT: plant height, GY: grain yield traits.

Table S5: Recombination rate and proportion of low and high recombination regions. Average recombination rate ("RecRate") and proportion of the physical ("Phys") and genetic ("Genetic) map in low ("LowRec", <0.5 cM / Mbp) and high ("HighRec", >0.5 cM / Mbp) recombination regions for each chromosomes. "Chr" indicates the chromosome. Physical and genetic size columns indicated the size of each chromosome in bp and cM, respectively.

Figure S13: Percentage of QTLs located in high (darkgrey) and low (lightgrey) recombinogenic regions. (a) male flowering time, (b) plant height and (c) grain yield.

Additional file 11 (.pdf):

Table S6: Description of inbred lines. Variety and accession along with the breeders, seeds providers and genetic groups obtained using ADMIXTURE for K=4 (Stiff Stalk, Iodent, Lancaster, Other).

Additional file 12 (.docx):

Table S7: Narrow sense heritability (h^2) and variance components (V_g, genetic variance; V_e, residual variance). The heritability and variance components were estimated for all traits (grain yield, male flowering time and plant height) using the R package Heritability [1].

Figure S14: Linkage disequilibrium based approach to delineate a physical window around each SNP, examplified with chromosome 3. Linkage disequilibrium (LD) windows were defined per chromosome for each SNP based on physical LD extent in low recombinogenic regions (left part) and based on genetic LD extent in high recombinogenic regions (right part). These LD windows were used (i) to group significant SNPs into QTLs when they overlapped, (ii) to estimate genome coverage to detect QTLs by GWAS considering region not covered by LD windows, (iii) identify putative underlying genes involved in trait variations.

Figure S15: QQ-plots representing observed *-log$_{10}$(p-value)* against expected *-log$_{10}$(p-value)* under null hypothesis (No association, black line). We tested association between 44,729 SNPs from the 50K array and the male flowering time trait in one environment (Gai12R) using different GWAS models, kinship estimators and programs. (A) Comparisons between statistical models: M1 is the model without correction (green dots), M2 takes into account the group structure (blue dots), M3 takes into account kinship (IBD: *K_freq*) between individuals (purple dots) and M4 takes into account both group structure and kinship (red dots). (B) Comparison between mixed models using different estimates (IBS and IBD, *K_freq*) of kinship. (C) Comparison of using or not Rincent *et al.* 2014 approach (using *K_Chr* vs *K_freq*). (D) Comparison between different informatics tools (EMMAX, ASReml, FasST-LMM) that perform GWAS.

Figure S16: Correlations between the GWAS results from the GBS genetic data using a kinship estimated from the PANZEA 50K array (x-axis) and a kinship estimated from the GBS (y-axis). The horizontal and vertical lines are the threshold *-log$_{10}$(p-value) = 5*. The correlations were done using the flowering time (DTA) and plant height (plantHT) traits and the two sites, two years and two treatments (Gai12R, Gai12W, Gai13R, Gai13W, Ner12R, Ner12W, Ner13R, Ner13W).