

Deciphering the genetic diversity of landraces with high-throughput SNP genotyping of DNA bulks: methodology and application to the maize 50k array Mariangela Arca

Mariangela Arca, Tristan Mary-Huard, Cyril Bauland, Brigitte Gouesnard, Aurélie Berard, Valérie Combes, Delphine Madur, Alain Charcosset, Stephane

Nicolas

▶ To cite this version:

Mariangela Arca, Tristan Mary-Huard, Cyril Bauland, Brigitte Gouesnard, Aurélie Berard, et al.. Deciphering the genetic diversity of landraces with high-throughput SNP genotyping of DNA bulks: methodology and application to the maize 50k array Mariangela Arca. 2020. hal-02965048

HAL Id: hal-02965048 https://hal.inrae.fr/hal-02965048

Preprint submitted on 12 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Deciphering the genetic diversity of landraces with high throughput SNP genotyping of DNA bulks: methodology and application to the maize 50k array

4	
5	Mariangela Arca ¹ , Tristan Mary-Huard ¹ , Brigitte Gouesnard ² , Aurélie Bérard ³ , Cyril
6	Bauland ¹ , Valérie Combes ¹ , Delphine Madur ¹ , Alain Charcosset ¹ , Stéphane D. Nicolas ¹
7	Author's affiliations :
8	1 Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE - Le Moulon, 91190,
9	Gif-sur-Yvette, France
10 11	2 Amélioration Génétique et Adaptation des Plantes méditéranéennes et tropicales, Univ Montpellier, CIRAD, INRAE, Institut Agro, F-34090 Montpellier, France
12	3 Université Paris-Saclay, INRAE, Etude du Polymorphisme des Génomes Végétaux,
13	91000, Evry-Courcouronnes, France
14	Corresponding authors: stephane.nicolas@inrae.fr

16 **ABSTRACT**

17 Genebanks harbor original landraces carrying many original favorable alleles for mitigating biotic and abiotic stresses. Their genetic diversity remains however poorly 18 characterized due to their large within genetic diversity. We developed a high-throughput, 19 20 cheap and labor saving DNA bulk approach based on SNP Illumina Infinium HD array to 21 genotype landraces. Samples were gathered for each landrace by mixing equal weights from 22 young leaves, from which DNA was extracted. We then estimated allelic frequencies in each 23 DNA bulk based on fluorescent intensity ratio (FIR) between two alleles at each SNP using a 24 two step-approach. We first tested either whether the DNA bulk was monomorphic or 25 polymorphic according to the two FIR distributions of individuals homozygous for allele A or 26 B, respectively. If the DNA bulk was polymorphic, we estimated its allelic frequency by using 27 a predictive equation calibrated on FIR from DNA bulks with known allelic frequencies. Our 28 approach: (i) gives accurate allelic frequency estimations that are highly reproducible across 29 laboratories, (ii) protects against false detection of allele fixation within landraces. We 30 estimated allelic frequencies of 23,412 SNPs in 156 landraces representing American and 31 European maize diversity. Modified Roger's genetic Distance between 156 landraces estimated 32 from 23,412 SNPs and 17 SSRs using the same DNA bulks were highly correlated, suggesting 33 that the ascertainment bias is low. Our approach is affordable, easy to implement and does not 34 require specific bioinformatics support and laboratory equipment, and therefore should be highly relevant for large-scale characterization of genebanks for a wide range of species. 35

Keywords: Landraces, DNA pooling, Genetic diversity, 50K Illumina Infinium HD Zea
mays L., Allelotyping , genebanks

38 INTRODUCTION

39 Genetic resources maintained in situ or ex situ in genebanks represent a vast reservoir 40 of traits/alleles for future genetic progress and an insurance against unforeseen threats to 41 agricultural production (Tanksley 1997; Hoisington et al. 1999; Kilian and Graner 2012; 42 McCouch et al. 2012). Due to their local adaptation to various agro-climatic conditions and 43 human uses, landraces are particularly relevant to address climate change and the requirements 44 of low input agriculture (Fernie et al. 2006; McCouch et al. 2012; Mascher et al. 2019). For instance, maize displays considerable genetic variability, but less than 5 % of this variability 45 has been exploited in elite breeding pools, according to (Hoisington et al. 1999). However, 46 47 landraces are used to a very limited extent, if any, in modern plant breeding programs, because they are poorly characterized, genetically heterogeneous and exhibit poor agronomic 48 49 performance compared to elite material (Kilian and Graner 2012; Strigens et al. 2013; Brauner et al. 2019; Mascher et al. 2019; Hölker et al. 2019). Use of molecular techniques for 50 51 characterizing genetic diversity of landraces and their relation with the elite germplasm is 52 essential for a better management and preservation of genetic resources and for a more efficient 53 use of these germplasms in breeding programs (Hoisington et al. 1999; Mascher et al. 2019).

54 The genetic diversity of landraces conserved ex situ or in situ has been extensively studied 55 using various types of molecular markers such as restriction fragment length polymorphism 56 (RFLP) or simple sequence repeat (SSR) in maize (Dubreuil and Charcosset 1998; Dubreuil et al. 1999; Rebourg et al. 1999, 2001; Gauthier et al. 2002; Rebourg et al. 2003; Reif et al. 2005b; 57 58 Vigouroux et al. 2005; Reif et al. 2005a; Camus-Kulandaivelu 2006; Dubreuil et al. 2006; 59 Eschholz et al. 2010; Mir et al. 2013), in Pearl Millet (Bhattacharjee et al. 2002), cabbage (Dias 60 et al. 1991; Divaret et al. 1999), Barley (Parzies et al. 2000; Backes et al. 2003; Hagenblad et 61 al. 2012), pea (Hagenblad et al. 2012), oat (Hagenblad et al. 2012), rice (Ford-Lloyd et al. 62 2001; McCouch et al. 2012), Alfalfa (Pupilli et al. 2000; Segovia-Lerma et al. 2003) and fonio 63 millet (Adoukonou-Sagbadja et al. 2007). SSRs have proven to be markers of choice for 64 analyzing diversity in different plant species and breeding research, because of their variability, 65 ease of use, accessibility of detection and reproducibility (Liu et al. 2003; Reif et al. 2006; Yang 66 et al. 2011). Nevertheless, the development of SSR markers requires a substantial investment of time and money. Allele coding is also difficult to standardize across genotyping platforms 67 68 and laboratories, a major drawback in a genetic resources characterization context. SNPs have 69 become the marker of choice for various crop species such as maize (Ganal et al. 2011), rice 70 (McCouch et al. 2010), barley (Moragues et al. 2010) and soybean (Lam et al. 2010). They are 71 the most abundant class of sequence variation in the genome, co-dominantly inherited, 72 genetically stable and appropriate to high-throughput automated analysis (Rafalski 2002). For instance, maize arrays with approx. 50,000 and 600,000 SNP markers are available since 2010 73 74 (Illumina MaizeSNP50 array, Ganal et al. 2011) and 2013 (600K Affymetrix Axiom, Unterseer 75 et al., 2013), respectively. SNP arrays may however lead to some ascertainment bias notably 76 when diversity analysis was performed on a plant germplasm distantly related from those that 77 have been used to discover SNP put on the array (Nielsen 2004; Clark et al. 2005; Hamblin 78 et al. 2007; Inghelandt et al. 2011; Frascaroli et al. 2013). Properties of SNP array regarding 79 diversity analysis have to be carefully investigated to evaluate ascertainment biais. For maize 50K Infinium SNP array, only "PZE" prefixed SNPs (so called later PZE SNPs in this study) 80 81 give consistent results for diversity analysis as compared with previous studies based on SSR markers and are therefore suitable for assessing genetic variability (Inghelandt et al. 2011; 82 83 Ganal et al. 2011; Bouchet et al. 2013; Frascaroli et al. 2013). 50K Infinium SNP array has 84 been used successfully to decipher genetic diversity of inbred lines (van Heerwaarden et al. 85 2011; Bouchet et al. 2013; Frascaroli et al. 2013; Rincent et al. 2014), landraces using either 86 doubled haploids (Strigens et al. 2013) or a single individual per accession (van Heerwaarden 87 et al. 2011; Arteaga et al. 2016), or teosinte with few individuals per accession (Aguirre-88 Liguori *et al.* 2017).

89 Due to high diversity within accessions, characterization of landraces from allogamous 90 species such as maize or alfalfa should be performed based on representative sets of individuals 91 (Reves-Valdés et al. 2013, Segovia-Lerma et al., 2002, Dubreuil and Charcosset., 1998). 92 Despite the recent technical advances, genotyping large numbers of individuals remains very 93 expensive for many research groups. To bring costs down, estimating allele frequencies from 94 pooled genomic DNA (also called "allelotyping") has been suggested as a convenient 95 alternative to individual genotyping using high-throughput SNP arrays (Sham et al. 2002; 96 Teumer et al. 2013) or using Next Generation Sequencing (Schlötterer et al. 2014). It was 97 successfully used to decipher global genetic diversity within maize landraces using RFLPs 98 (Dubreuil and Charcosset 1998; Dubreuil et al. 1999; Rebourg et al. 2001, 2003; Gauthier et 99 al. 2002) and SSR markers (Reif et al. 2005a; Camus-Kulandaivelu 2006; Dubreuil et al. 2006; 100 Yao et al. 2007; Mir et al. 2013). Various methods for estimating gene frequencies in pooled 101 DNA have been developed for RFLP (Dubreuil and Charcosset 1998), SSR (LeDuc et al. 1995; 102 Perlin et al. 1995; Daniels et al. 1998; Lipkin et al. 1998; Breen et al. 1999) and SNP marker 103 arrays in human and animal species (Hoogendoorn et al. 2000; Craig et al. 2005; Brohede 2005;

104 Teumer et al. 2013; Gautier et al. 2013). These methods have been successfully used to detect 105 QTL (Lipkin et al. 1998), to decipher genetic diversity (Segovia-Lerma et al. 2003; Dubreuil 106 et al. 2006; Pervaiz et al. 2010; Johnston et al. 2013; Ozerov et al. 2013), to perform genome 107 wide association studies (Barcellos et al. 1997; Sham et al. 2002; Baum et al. 2007), to identify 108 selective sweep (Elferink et al. 2012) or to identify causal mutation in tilling bank (Abe et al. 109 2012). Genotyping DNA bulks of individuals from landraces with SNP arrays could therefore 110 be interesting to characterize and manage genetic diversity in plant germplasm. SNP arrays could be notably a valuable tool to identify selective sweep between landraces depending on 111 112 their origin, to manage plant germplasm collection at worldwide level (e.g. identify duplicate), 113 to identify landraces poorly used so far in breeding programs or to identify genomic regions 114 where diversity has been lost during the transition from landraces to inbred lines (Arca et al., in 115 prep).

116 In this study, we developed a new DNA bulk method to estimate allelic frequencies at SNPs 117 based on Fluorescent Intensity data produced by the maize 50K Illumina SNP array (Ganal et 118 al. 2011). Contrary to previous methods that have been mostly developed for QTL detection 119 purposes, our approach is dedicated to genome-wide diversity analysis in plant germplasm since 120 it protects against false detection of alleles. Additionally, calibration of equations for predicting 121 allelic frequencies of DNA bulks for each SNP is based on controlled pools with variable allelic 122 frequencies rather than only heterozygous genotypes as in previous methods (Hoogendoorn et 123 al. 2000; Brohede 2005; Peiris et al. 2011; Teumer et al. 2013). As a proof of concept, we 124 applied our new approach to maize by estimating allelic frequencies of 23,412 SNPs in 156 125 maize landraces representative of European and American diversity present in genebanks 126 (Arca et al., in prep). To our knowledge, it is the first time that a DNA bulk approach was used 127 on 50K maize high-throughput SNP array to study genetic diversity within maize landraces 128 germplasm.

129 *RESULTS*

We developed a new method to estimate allelic frequencies of SNPs within pools of individuals using the fluorescent intensity ratio (FIR) between A and B alleles from Illumina MaizeSNP50 array. Briefly, allelic frequencies at SNPs belonging to MaizeSNP50 array were estimated within 156 maize landraces by pooling randomly 15 individuals per population and by calibrating a predictive two-step model (Figure 1). We considered only the subsample of 32,788 prefixed PZE markers (so called PZE SNPs) that have proven suitable for diversity analyses (Ganal *et al.* 2011). Among these SNPs, we selected 23,412 SNPs that passed weighted
deviation (wd) quality criteria (wd>50). This removed SNPs for which estimated allelic
frequency deviated strongly from expected allelic frequency (Figure. S1 A, B, C, D, E, F and
G for the threshold choice and validation).

140 Accuracy of allelic frequency prediction and detection of allele fixation

141 In order to prevent erroneous detection of alleles within landraces, we first tested for each 142 landrace whether allele A or allele B was fixed at a given SNP locus (Figure 1). We tested for 143 each SNP whether the FIR of the landrace was included within one the two Gaussian 144 distributions drawn from mean and variance of FIR of genotypes AA and BB within the inbred 145 line panel (Figure 1). For landraces that were considered polymorphic after this first step (allele 146 fixation rejected for both alleles), we estimated allelic frequency based on FIR by using a unique 147 logistic regression model for the 23,412 SNPs, calibrated with a sample of 1,000 SNPs (Figure 148 1). Parameters of the logistic model were adjusted on these 1,000 SNPs using FIR of two series 149 of controlled pools whose allelic frequencies were known (Figure S2). We obtained these pools 150 by mixing various proportion of two series of three inbred lines with known genotypes (Table 151 1). The1,000 SNPs were selected to maximize the allelic frequency range within controlled 152 pools (Table 1). The logistic regression was calibrated on 1,000 SNPs taken together rather than 153 for each SNP individually to avoid the ascertainment bias that would be generated by selecting 154 only SNPs polymorphic in the controlled pools (Figure S3) and to reduce loss of accuracy in 155 prediction for SNPs displaying limited allelic frequency range in two controlled pools (Figure 156 S4). To investigate the loss of accuracy of the prediction curve due to a reduction in allelic 157 frequency ranges in controlled pool, we progressively removed at random from one to 15 158 samples from the calibration set of the 1000 above described SNPs. The mean absolute error 159 (MAE) between 1000 replications increased significantly from 4.14 % to 8.54 % when 160 removing more samples (Table 2). For comparison, MAE was 7.19 % using a cross-validation 161 approach in which the predictive equation was calibrated with a random subsample of 800 out 162 of 1000 SNPs, and then applied to estimate allelic frequencies for the remaining 200 SNPs 163 (Table S1). Calibrating the logistic regression between FIR and allelic frequency in controlled 164 pool based on 1000 SNPs therefore appears well adapted to prevent ascertainment bias while 165 increasing globally prediction accuracy (Figure S4). Finally, we observed that MAE was higher 166 for balanced allelic frequencies than for almost fixed allelic frequencies (Figure 2 and Tables 167 S2). Accordingly, the dispersion of predicted frequencies were larger for expected allelic 168 frequencies near 0.5 than for fixed alleles (Table S2).

169 *Reproducibility of frequency across laboratories and samples*

170 We evaluated the reproducibility of the method across laboratories by comparing FIR of 171 one series of controlled pools from two different laboratories using all PZE SNPs or 23,412 172 SNPs selected using wd criterion (Figure 3). The correlation between the two different laboratories for controlled pools was very high ($r^2 > 0.98$) whether we selected the SNPs based 173 174 on *wd* criterion or not. Beyond reproducibility across laboratories, the precision of frequency 175 estimation depends on the sampling of individuals within landraces (Table 3). The precision of 176 frequency estimation was addressed both by numerical calculation and by the independent 177 sampling of 15 different individuals (30 different gametes) within 10 landraces (biological 178 replicate). For both numerical calculations and biological replicates, the sampling error was 179 higher for loci with balanced allelic frequencies than for loci that are close to fixation (Table 3, 180 Figure 4). Sampling error also decreased as the number of sampled individuals increased (Table 181 3). Considering a true frequency of 50% within landraces, we expect that 95% of frequency 182 estimates lie between 31.30% and 68.70% when sampling 15 individuals per landrace and 42.9 183 to 57.13% when sampling 100 individuals per landrace (Table 3). Considering biological 184 replicates, allelic frequencies of the two biological replicates over 23,412 SNPs were highly 185 correlated except for population Pol3 (Table S3). When excluding Pol3, 94.5% of points were 186 located within the 95% confidence limits accounting for the effect of sampling alone, 187 suggesting that the error inherent to the frequency estimation for DNA pools was very low 188 compared to the sampling error (Figure 4). Over nine populations with replicates (excluding 189 Pol3), we observed only four situations among 23,412 loci where two different alleles were 190 fixed in the two replicates (Figure 4). Loci for which an allele was fixed in one replicate was 191 either fixed or displayed a high frequency (above 88%) for the same allele in the other replicate 192 in 98% of cases. Moreover, we estimated pairwise roger's genetic distance (MRD) based on 193 23,412 SNPs between the two independent pools from the same landraces. Excluding 194 Population Pol3 (MRD = 0.208), this distance ranged from 0.087 to 0,120 (Table S3). These 195 values provide a reference to decide whether two populations can be considered different or 196 not.

Effect of SNP number and wd on the relationship of genetic distance estimated with SNP and SSR

Finally, we evaluated the possible ascertainment bias due to SNP selection with our filtering based on *wd* criterion. MRD obtained with 17 SSR markers (MRD_{SSR}) and MRD based on

- 201 different set of SNP markers (MRD_{SNP}) were highly correlated (Figure 5), indicating a low
- ascertainment bias. The selection of SNPs based on *wd* quality criterion strongly increased the
- 203 coefficient of determination (r^2) between MRD_{SNP} and MRD_{SSR}, from 0.587 to 0.639 (Figure
- 204 S6). We attempted to define the minimal SNP number required to correctly describe the
- relationship between maize landraces. While increasing the number of SNPs from 500 to 2500
- slightly increased r^2 between MRD_{SNP} and MRD_{SSR} from 0.606 to 0.638 (Figure S6 D,E,F), we
- 207 observed no further increase beyond 2500 SNPs (Figure S6 A, B, C) suggesting that 2,500
- 208 SNPs are enough to obtain a correct picture of landrace relationships.

209 **DISCUSSION**

210 A molecular approach for diversity analysis of landraces needs to answer several criteria 211 (i) an accurate estimation of allelic frequency in each population, (ii) a robust and reproducible 212 measurement of allelic frequency across laboratories in order to facilitate comparison of genetic 213 diversity of accessions across genebanks, (iii) a reliable estimate of genetic distance between 214 landraces with no or little ascertainment bias (iv) an affordable, high-throughput and labor 215 efficient method, due to both strong financial and human constraints in plant genebanks. Four 216 main sources of errors affect the accuracy of allelic frequency estimation of a locus in a 217 population using a DNA pooling approach: (i) the sampling of individuals (so called "sampling" 218 errors), (ii) the procedure to mix DNA from individuals (so called "DNA mixing" errors) (iii) 219 the imprecision of quantitative measurement used by the model for the prediction (so called 220 "experimental" errors) and (iv) the predictive equation used to predict allelic frequency in a 221 population (so called "approximation" errors).

222 A two-step model to protect against erroneous detection of 223 polymorphism and predict accurately allelic frequencies in DNA bulk

224 Approximation errors due to predictive equation depend on (i) the model used to predict 225 allelic frequencies and (ii) the set of individuals and SNPs used to calibrate the predictive 226 equation. In this study, we used a two-step modeling using inbred lines and controlled pools as 227 sets of calibration to test for polymorphism and then predict allelic frequency for polymorphic 228 markers. Detection of allele fixation in a population is an important issue for deciphering 229 and managing genetic diversity within plant and animal germplasm. We used two Student 230 tests based on fluorescent intensity ratio (FIR) distribution of lines homozygous for allele A and B to determine polymorphism of a SNP in a given landrace (Figure 1). In this first 231 232 step, we preferred a method based on FIR distribution rather than the clustering 233 approach implemented in genome studio because it is possible to control the type I risk 234 of false allele detection (at 5% in our study). Using this two-step approach reduces 235 strongly the erroneous detection of polymorphisms in a population compared to previous 236 methods: MAE for fixed locus <0.1% in our approach (Table S2) vs \sim 2-3% using PPC 237 method (Brohede 2005) or ~2-8% using different k correction from (Peiris *et al.* 2011). 238 This is not surprising as previous methods focused on an accurate estimation of the 239 difference in allele frequencies between DNA bulks of individuals contrasted for a

quantitative trait of interest (Sham *et al.* 2002; Craig *et al.* 2005; Kirov *et al.* 2006; Teumer *et al.* 2013) and did not focus specifically on protecting again false detection of alleles.

242 For loci that were detected as polymorphic, we predicted allelic frequencies from 243 FIR in landrace DNA pools by using a unique logistic regression for 23,412 SNPs passing 244 *wd* quality criterion. The relationship between FIR and allelic frequency was modelled 245 using a quasi-logistic regression for different reasons. First, the logistic function ensures 246 that the predicted frequencies take value in (0,1), a property that is not satisfied by 247 polynomial regression (PPC) or tan transformation (Brohede 2005; Teumer et al. 2013). 248 Second, one could observe that the relationship between the fluorescent intensity ratio 249 and allelic frequencies within controlled pools was not linear (Figure S2).

250 This two-step approach led to a low global error rate in allelic frequency prediction (MAE = 3% for polymorphic and monomorphic loci considered jointly; Figure 2, Table S2). 251 252 It is comparable to the most accurate previous pooling DNA methods for SNP array that used a specific model for each SNP: (i) MAE ranging from 3 to 8 % (Peiris et al., 2011) or 253 254 5-8% (Brohede et al., 2005) depending of k-correction applied (ii) MAE \sim 3% for PPC correction (Brohede et al., 2005; Teumer et al., 2013) (iii) MAE $\sim 1\%$ for tan-correction 255 256 (Teumer et al., 2013). Several factors can explain this relative good global accuracy of our 257 approach. First, almost half of the loci were fixed on average in each landrace, which 258 contributed positively to global accuracy since our method over-performed previous 259 methods for fixed locus (see above). Second, wd quality criterion removed SNPs for which 260 allelic frequencies were poorly predicted using FIR. We observed indeed that increasing 261 the threshold for *wd* quality criterion led to a global increase in accuracy at both steps 262 (Figure S1). While 90% of SNPs have a MAE<2% for wd criterion >10, only 50% of SNPs 263 have a MAE<2% for wd criterion <10. Taking into account differential hybridization by 264 using a specific logistic regression for each SNP could be a promising way to further 265 improve the accuracy of allelic frequencies prediction, notably for balanced allelic 266 frequencies (Brohede et al., 2005, Peiris et al., 2011, Teumer et al., 2013). To limit possible 267 ascertainment bias and errors in allelic frequency estimation, it requires however to 268 genotype additional series of controlled pools for SNPs for which current controlled pools 269 were monomorphic or have a limited range of allelic frequency (Figure S3 and S4).

270 To estimate the parameters of the logistic regression, we used two series of controlled 271 pools rather than heterozygous individuals for both technical and practical reasons. 272 Controlled pools cover more homogenously the frequency variation range than 273 heterozygous and homozygous individuals only, which therefore limits the risk of 274 inaccurate estimation of logistic model parameters. Different studies showed that 275 accuracy of allelic frequency estimation strongly depends on accuracy of FIR estimation 276 for heterozygous individual and therefore the number of heterozygous individuals (Le 277 Hellard et al. 2002; Simpson 2005; Jawaid and Sham 2009). Between 8 and 16 278 heterozygous individuals are recommended to correctly estimate FIR mean for 279 heterozygous individuals, depending on FIR variance (Le Hellard et al. 2002). In maize, 280 we can obtain heterozygote genotypes either by crossing inbred lines to produce F1 281 hybrids, by planting seeds from maize landraces, or by using residual heterozygosity of 282 inbred lines. Using residual heterozygosity to calibrate model is not possible since half 283 SNPs show no heterozygous genotype in the 327 inbred lines of our study. Obtaining at least 16 heterozygous individuals for each SNP therefore requires to genotype a few 284 285 dozens of F1 hybrids or individuals from landraces considering that expected 286 heterozygosity in a landrace is comprised between 3 and 28% (Arca et al., in prep). This 287 represents additional costs since maize researchers and breeders genotyped 288 preferentially inbred lines to access directly haplotypes without phasing and because 289 genotypes of F1 hybrids can be deduced of that of their parental inbred lines. Beyond 290 allogamous species as maize, genotyping heterozygous individuals could be time 291 demanding and very costly in autogamous cultivated plant species for which genotyped 292 individuals are mostly homozygotes (wheat, tomato, rapeseed). On the contrary, one can 293 easily produce controlled pools whatever the reproductive system, either by mixing DNA 294 or equal mass of plant materials, which allows producing a wide range of allelic 295 frequencies.

296 *Effect of DNA mixing procedure on accuracy allelic frequency estimation*

There are two main errors coming from DNA mixing procedure: (i) the "sampling error" that is directly connected to the number of individuals sampled in each population (Table 3), and (ii) the "bulking error" associated with the laboratory procedure to mix equal DNA amounts of sampled individuals. 301 We evaluated sampling and bulking errors by comparing 10 independent biological 302 replicates from 10 different landraces obtained by independently sampling and mixing 303 equal leaf areas of young leaves of 15 individuals. Allelic frequencies estimated for both 304 biological replicates from a same landrace were highly correlated. Excluding Pol3, 94.5% 305 of difference of allelic frequencies between replicates was of included within 95% confidence limits originated from sampling effect only Figure 4). This suggests that the 306 307 "bulking error" is low compared to the "sampling error". Consistently, Dubreuil et al., 308 (1999) observed a low "bulking error" for RFLP markers using the same DNA pooling 309 method, with a coefficient of determination of 0.99 between allelic frequencies based on 310 individual genotyping of plants and those predicted using DNA bulks. Several studies also 311 showed that the effect of bulking errors on allelic frequencies measured by comparing 312 DNA pool and individual genotyping of plant of this DNA pool is very low compared with 313 other sources of errors (Le Hellard et al. 2002; Jawaid and Sham 2009). Additionally, the 314 mixing procedure starting from leaf samples strongly reduced the number of DNA 315 extractions for each DNA bulk as compared to first extracting DNA from each individual, 316 and then mixing by pipetting each DNA samples to obtain an equimolar DNA mix ("post-317 extraction" approach). Since the cost of DNA extraction becomes non-negligible when the 318 number of individual increases, mixing plant material based on their mass before 319 extraction is highly relevant to save time and money. This can be done without losing 320 accuracy as shown in this study for SNP array and previously for RFLP by Dubreuil et al., 321 (1999).

322 We highlighted the critical importance of the number of individuals sampled per 323 landrace on allelic frequency estimation (Table 3). By using DNA pooling, accuracy can be 324 gained with very little additional cost by increasing number of sampled individuals. 325 Whereas a high accuracy of allelic frequency estimation within landraces is required to 326 scan genome for selective sweeps, it is less important to estimate global genetic distance, 327 due to the large number of SNPs analyzed. Sampling fifteen plants per population (30 328 gametes) appears convenient to obtain an accurate estimation of frequencies in a 329 population and analyze genetic diversity (Reyes-Valdés et al. 2013).

330 A low ascertainment bias to estimate genetic distance between landraces

There are two possible sources of ascertainment bias using a DNA pooling approach on a SNP array. The first one relates to the design of array because the set of lines to discover SNPs may not well represent genetic diversity and a threshold in allelic frequency was possibly applied to select SNPs. The second one relates to the selection of a subset of SNPs from the array regarding the genetic diversity of samples in calibration set used to predict allelic frequencies.

To avoid risk of ascertainment bias due to selection of markers genotyped by the 337 338 array, the logistic regression model was adjusted on 1,000 SNPs with the largest allelic 339 frequency range rather than for each of the 23,412 PZE SNPs individually. Using a specific 340 model for each SNP would indeed conduct to discard markers monomorphic in controlled 341 pools and therefore select only markers polymorphic between parents of controlled pool. 342 Note that the same issue would be raised by using heterozygous individuals since 8 to 16 heterozygotes were recommended to adjust a logistic regression. Using heterozygous 343 344 individuals and SNP specific equations could lead to systematically counter-select SNPs 345 with low diversity. It could also lead to systematically remove SNPs that are differentially 346 fixed between isolated genetic groups, because no or very few heterozygote individuals 347 are available.

348 We also evaluated ascertainment bias by comparing Modified Roger's Distance (MRD) 349 between the 156 landraces obtained using SNPs (MRD_{SNP}) and SSRs (MRD_{SSR}) (Camus-Kulandaivelu 2006; Mir et al. 2013), which display no or limited ascertainment bias. 350 351 MRD_{SNP} was highly correlated with MRD_{SSR} ($r^2 = 0.64$; Figure 5). This correlation is high 352 considering that SSR and SNP markers evolve very differently (mutation rate higher for 353 SSRs than SNPs, multiallelic vs biallelic), that the number of SSR markers used to estimate 354 genetic distance is low and that errors in allelic frequency prediction occur for both SNPs 355 and SSRs. For comparison, correlation was lower than between Identity By State estimated with 94 SSRs and 30k SNPs in a diversity panel of 337 inbred lines ($r_2 = 0.41$) 356 357 , although very few genotyping errors are expected in inbred lines (Bouchet et al. 2013). 358 Using the *wd* criterion significantly increased the correlation between MRD_{SNP} and MRD_{SSR} 359 markers for 156 landraces (Figure S5). It suggests that the wd criterion removes SNP 360 markers that blurred the relationships between landraces. We can therefore define a subset of 23,412 SNPs to analyze global genetic diversity in landraces. This is in 361

agreement with previous studies in inbred lines showing that PZE SNPs are suitable to
analyze the genetic diversity in inbred lines (Inghelandt *et al.* 2011; Ganal *et al.* 2011;
Bouchet *et al.* 2013; Frascaroli *et al.* 2013). These studies showed that diversity analysis
based on PZE SNPs give consistent results with previous studies based on SSR markers
(Inghelandt *et al.* 2011; Bouchet *et al.* 2013; Frascaroli *et al.* 2013).

The DNA pooled-sampling approach therefore provides a reliable picture of the genetic relatedness among populations that display a large range of genetic divergence and opens a way to explore genome-wide diversity along the genome.

An affordable, high-throughput, labor-efficient and robust method compared to SSR / RFLP markers and sequencing approaches

372 Using SNP arrays instead of SSR/RFLP marker systems or sequencing approaches 373 has several advantages. First, SNP genotyping using arrays is very affordable compared 374 to SSR/RFLP or resequencing approaches because it is highly automatable, high-375 throughput, labor-efficient and cost effective (currently 30-80€ / individual depending of 376 array). Obtaining accurate estimations of allelic frequencies using a whole genome 377 sequencing (WGS) approach requires high depth and coverage for each individual 378 because of the need of counting reads (Schlötterer et al. 2014). To estimate allelic 379 frequency in DNA bulks, WGS remains costly compared to SNP arrays for large and 380 complex genomes of plant species as maize. Different sequencing approaches based either 381 on restriction enzyme or sequence capture make it possible to target some genomic 382 regions and multiplex individuals, reducing the cost of library preparation and 383 sequencing while increasing the depth for the selected regions (Glaubitz et al. 2014). 384 However, these sequencing approaches remain more expensive than SNP arrays and 385 require laboratory equipment to prepare DNA libraries and strong bioinformatics skills 386 to analyze sequencing data. These skills are not always available in all genebanks. With 387 the maize 50K array, FIR measurement used to predict allelic frequencies were highly reproducible both across laboratories and batches (r2 = 0.987; Figure 3). We can 388 389 therefore consistently predict allelic frequencies using 50K array in new DNA pools 390 genotyped in other laboratories, by applying the same parameters of presence /absence 391 test and logistic regression as in this study. This will greatly facilitate the comparison of 392 accessions across collections and laboratories. This is a strong advantage over SSRs for

393 which a strong laboratory effect has been observed for the definition of alleles, leading to 394 difficulties for comparing genetic diversity across seedbanks and laboratories (Mir et al. 395 2013). Similarly, one can expect some laboratory effect for sequencing approaches due to 396 preparation of library and bioinformatics analysis. However, there is some disadvantage 397 to use SNP arrays instead of SSR markers or sequencing approach. First, SNP marker are 398 bi-allelic whereas SSRs are multi-allelic. At a constant number of markers, using SNPs 399 rather than SSRs therefore leads to less discriminative power (Laval et al. 2002; Hamblin 400 et al. 2007). This disadvantage is largely compensated by the higher number of SNPs and 401 the fact that SNPs are more frequent and more regularly spread along the genome than 402 SSR/RFLP, allowing genome wide diversity analyses. Second, contrary to SSR / RFLP 403 markers and sequencing approach, SNP array does not allow one to discover new 404 polymorphisms, which may lead to ascertainment bias for diversity analysis of new 405 genetic groups (Nielsen 2004; Clark et al. 2005; Hamblin et al. 2007; Inghelandt et al. 406 2011; Frascaroli et al. 2013). Comparison with SSRs results showed that PZE SNPs 407 provide reliable genetic distances between landraces, suggesting a low ascertainment 408 bias for a global portrayal of genetic diversity (see above). Sequencing techniques may be 409 interesting in a second step to identify, among preselected accessions, those which show 410 an enrichment in new alleles.

411 The number of SNPs affects the estimates of relationship between landraces and 412 population structure (Moragues et al. 2010). In our study, the correlation coefficient 413 between MRD_{SNP} and MRD_{SSR} increased with increasing number of SNPs and reached a 414 plateau for 2,500 SNPs (Figure S6). This suggests that increasing the number of SNPs 415 above 2,500 does not provide further improvement in precision to estimate relationships 416 between landraces as compared to 17 SSRs. Our approach could therefore be made further 417 cost efficient by selecting less loci for studying global genetic relationships and genetic 418 diversity. For maize, a customizable 15K Illumina genotyping array has been developed that 419 includes 3,000 PZE SNPs selected for studying essential derivation (Rousselle et al. 2015) and 420 12,000 others selected for genetic applications such as genomic selection. Alternatively, the 421 same approach could be applied to other genotyping arrays with higher density as the 600K 422 Affymetrix Axiom Array (Unterseer et al., 2013) to gain precision in detection of selective 423 footprints.

424 **CONCLUSION**

425 The DNA pooling approach we propose overcomes specific issues for genetic diversity 426 analysis and plant germplasm management purposes that were not or partially addressed by 427 previous methods which were mostly focused on QTL analysis and genome wide association 428 studies (Hoogendoorn et al. 2000; Craig et al. 2005; Brohede 2005; Teumer et al. 2013). As 429 proof of concept, we used the DNA pooling approach to estimate allelic frequencies in maize 430 landraces in order to identify original maize landraces in germplasm for pre-breeding 431 purposes and selective footprints between geographic and/or admixture groups of 432 landraces cultivated in contrasted agro-climatic conditions (Arca et al., in prep). Our 433 approach could be very interesting for studying plant germplasm since time, money and 434 molecular skills can be limiting factors to study and compare large collections of landraces 435 maintained in seedbank (Mir et al. 2013). Applications could be expanded to QTL 436 identification in pools (Gallais et al. 2007), detecting signatures of selection in multi-generation 437 experiments, or detection of illegitimate seed-lots during multiplication in genebanks. The DNA 438 pooling approach could be easily applied to decipher organization of genetic diversity in 439 other plant germplasm since Infinium Illumina HD array have been developed for several 440 cultivated plant species, including soybean, grapevine, potato, sweet cherry, tomato, 441 sunflower, wheat, oat, brassica crops but also animal species.

442 MATERIALS AND METHODS

443 Plant material

444 Landraces

A total of 156 landrace populations (Table S4) were sampled among a panel of 413
landraces capturing a large proportion of European and American diversity and analyzed in
previous studies using RFLP (Dubreuil and Charcosset 1998; Rebourg *et al.* 1999, 2001, 2003;
Gauthier *et al.* 2002) and SSR markers (Camus-Kulandaivelu 2006; Dubreuil *et al.* 2006; Mir *et al.* 2013).

Each population were represented by a bulk of DNA from 15 individual plants, mixed in equal amounts as described in Reif *et al.* (2005) and Dubreuil *et al.* (2006). In order to analyze the effect of individual sampling on allelic frequency estimation (see below), ten populationswere represented by two DNA bulks of 15 plants sampled independently (Table 3).

454 Controlled DNA Pools

To calibrate a prediction model for SNP allelic frequencies in populations, we considered two series of nine controlled pools derived from the mixing of two sets of three parental inbred lines: EP1 - F2 - LO3 (European Flint inbred lines) and NYS302- EA1433 - M37W (Tropical inbred lines).

459 For each set of three parental lines, we prepared nine controlled pools by varying the 460 proportion of each line in the mix (Table 1), measured as the number of leaf disks with equal 461 size according to Dubreuil et al., (1999). The proportion of lines 2 and 3 (EA1433 and M37W 462 or F2 and LO3) varies similarly whereas line 1 (EP1 or NYS302) varies inversely. The genotype 463 of the inbred lines and the proportion of each inbred line in each pool give the expected allelic 464 frequencies as shown in Table 1. Combination of genotypes in parental lines can conduct either 465 to monomorphic or polymorphic controlled pools if the genotypes of 3 parental lines are the 466 same or not, respectively. If we exclude monomorphic controlled pools and heterozygote SNPs 467 in parental lines, these different combinations conduct to four different polymorphic 468 configurations in the 9 controlled pools, corresponding to four ranges of allelic frequencies: 1-469 33% (R1), 33-50% (R2), 51-67% (R3), 67-99% (R4), (Table 1). Combination of R1 and R4 470 configurations in two series of controlled pools displayed the largest allelic frequencies range 471 (1% to 99%) while combination of R2 and R3 displayed a more reduced allelic frequency range 472 (33% to 67%).

473 Inbred lines

To test for allele fixation within landraces, we used a panel of 333 inbred maize lines representing the worldwide diversity well characterized in previous studies (Camus-Kulandaivelu 2006; Bouchet *et al.* 2013) (Table S5). This panel includes the six inbred lines used to build two series of controlled pools.

478 Genotyping

We used the 50K Illumina Infinium HD array (Ganal *et al.* 2011) to genotype (i) 166 DNA
bulks representing 156 landraces (ii) 18 DNA bulks representing 2 series of controlled DNA
pools (iii) 333 inbred lines. 50K genotyping was performed according to the manufacturer's
instructions using the MaizeSNP50 array (IlluminaInc, San Diego, CA). The genotype results

were produced with GenomeStudio Genotyping Module software (v2010.2, IlluminaInc) using
the cluster file MaizeSNP50_B.egt available from Illumina. The array contains 49,585 SNPs
passing quality criteria defined in (Ganal *et al.* 2011).

We also used 17 SSRs genotyping data from 145 and 11 landraces analyzed by Camus etal. (2006) and Mir et al. (2013), respectively.

488 Measurement variable: fluorescence intensities ratio

The MaizeSNP50 array has been developed into allele-specific single base extension using two colors labeling with the Cy3 and Cy5 fluorescent dyes. The fluorescent signal on each spot is digitized using GenomeStudio software. Data consist of two normalized intensity values (x, y) for each SNP, with one intensity for each of the fluorescent dyes associated with the two alleles of the SNP. The alleles measured by the x intensity value (Cy5 dye) are arbitrary, with respect to haplotypes, are called the A alleles, whereas the alleles measured by the y intensity value (Cy3 dye) are called the B alleles.

We assumed that the strength of the fluorescent signal of each spot is representative of the amount of labeled probe associated with that spot. The amount of labeled probes at each spot relies upon the frequency of the corresponding alleles of PCR product immobilized on it. Based on this assumption, the fluorescent intensity ratio (FIR) of each spot (y/(x+y)) can be employed to estimate the allele frequency of DNA bulk immobilized on it.

501 To test the reproducibility of the measurement the controlled pool of European lines was 502 genotyped twice in two platforms, at CNG Genotyping National Center, Evry 91, France, and 503 at Trait Genetics.

504 SNP filtering and quality control

505 For the purpose of this study, we used only the subset of 32,788 markers contributed by the 506 Panzea project (http://www.panzea.org/), so called PZE SNPs, developed on the basis of US 507 NAM founders (Zhao 2006). These SNPs represent a comprehensive sample of the maize 508 germplasm and are therefore suitable for diversity analysis (Ganal *et al.* 2011).

509 The following equation (1) was then used to create a rank score (weighted deviation, wd) 510 for each SNP in order to identify and remove those of poor quality,

511
$$wd = \frac{|\mu AA - \mu BB|}{\sqrt{\frac{NAA \cdot \sigma AA^2 + NBB \cdot \sigma BB^2}{NAA + NBB}}} (1)$$

512 where μ_{AA} and σ_{AA} and μ_{BB} and σ_{BB} are the mean and the standard deviation for the 513 fluorescence intensity ratios of AA and BB genotypes for the 327 inbred lines panel and N_{AA} 514 and N_{BB} is the number of inbred lines with genotype AA or genotype BB respectively. To avoid 515 selection bias, loci which were monomorphic within the reference inbred lines population were 516 selected using the *wd* equations (1), assuming $\mu_{AA}=0$ and $\sigma_{AA}=0$ for monomorphic BB SNPs 517 or assuming $\mu_{BB}=1$ and $\sigma_{BB}=0$ for monomorphic AA SNPs.

This criterion removes from analysis those SNPs for which distributions of fluorescence signal ratios for AA and BB genotypes of 327 inbred lines panel overlap or have large variances. To analyze genetic diversity, we first selected 23,656 with *wd* above 50 among 32,788 PZE SNPs. This threshold removed SNPs displaying high error rate in allelic frequency prediction (Figure S1). In addition, we removed 244 SNPs that were heterozygous in one of parental lines of controlled pools and that displayed high error rate in allelic frequency prediction (data not shown).

525 Alleles detection and allele frequency estimation

Allele frequency estimation within DNA pools was implemented as a two-step process. We first determined the fixation of alleles A and/or B by comparing the fluorescent ratio of DNA pools at a given SNP locus with the distribution of the fluorescent signal of inbred lines (see above) which have AA or BB genotypes at the same locus. We assumed Gaussian distributions for the fluorescent intensities and tested for fixation using a Student's t-tests with a 5% type I nominal level.

532 In second step, for each SNP for which alleles A and B were both declared present, the 533 allelic frequency fB of allele B was inferred using the following generalized linear model:

534
$$g(fB) = \propto +\beta \frac{y}{x+y}(2)$$

where *x* and *y* are the fluorescent intensities at SNP for alleles A and B respectively, α and β are the parameters of a logistic curve, calibrated on fluorescent ratio data from controlled pools for 1000 SNPs and ε_i is a noise term. As allele B frequency is a binomial variable, GLM was set with a logit link function (R, version 3.0.3).

539 The calibration sample of 1,000 SNPs consists in 250 randomly selected SNPs for each 540 possible configuration (R1, R2, R3, R4 defined in Table 1). It was preferred to a calibration 541 sample of all SNPs or to a specific prediction curve for each SNPs, in order to have a 542 homogeneous distribution of observations into each class of expected frequency. Calibrating the model for each SNP would lead to high error in allelic frequency prediction, notably for monomorphic controlled pools as exemplified by Figure S3 and S4. Calibrating model for all SNPs would give strong weight to fixed allele in calibration due to large number of monomorphic controlled pools that are homozygous either for allele A or B.

547 Accuracy of allelic frequency estimation

We assessed the accuracy of allele frequency estimates from pooled DNA samples by calculating the absolute difference between allelic frequencies of the B allele predicted by our two-step model and those expected for controlled pools from the genotype of their six parental lines. We obtained expected allelic frequencies for two series of controlled pools by weighting the allelic frequency of each parental line (0 or 1) by their relative mass in the mix (Table 1). We obtained genotypes of inbred lines from clustering by genome studio. This absolute difference was averaged over SNPs and samples in order to obtain mean absolute error (MAE).

555 We first evaluated the mean absolute error for 23,412 SNPs in the two series of controlled 556 pools (Table S2, Figure 2). In order to estimate the effect of the calibration set of individuals 557 and SNPs on the accuracy of allelic frequency prediction, we applied two cross-validation 558 approaches on the 1000 SNPs and the two series of controlled pools and six parental inbred 559 lines (24 samples) used to calibrate parameters of the common logistic regression. In order to 560 evaluate the effect of SNP calibration set (Table S1), we repeated five time a K-fold approach 561 in which 1000 SNPs were split randomly in a training set of 800 SNPs on which we calibrated 562 our two-step model and a validation set of 200 SNPs on which we predicted allelic frequency 563 using this model in same two series controlled pools and estimated MAE. In order to evaluate 564 the effect calibration samples (Table 2), we repeated 1000 times a K-fold approach on 1000 565 SNPs in which 1, 3, 5, 8, 10, 15 samples among 18 from controlled pools were randomly removed from the calibration set. We used the remaining samples to estimate parameters of the 566 567 logistic regression, and then predicted allelic frequencies using this predictive equation in these 568 K removed samples (Table 2).

To estimate sampling error (Table 3), we estimated the 95% confidence interval of the allelic frequency in the population considering various observed allelic frequency obtained by sampling either 15, 30, 100 or 200 individuals from this population. To obtain the lower and upper bound of the 95% confidence interval for allelic frequency in the population, we considered the binomial probability to obtain various number of allele B in 15, 30, 100, 200 individuals (estimated allelic frequencies) from a population (true allelic frequencies) by using 575 binom.confint function implemented in R package "binom". We used the following parameters: 576 binom.confint(x = number of alleles observed, n = 2*number of individuals, conf.level=95%, 577 methods = exact) with x = number of successes and n = number of trial in the binomial 578 experiment.

579 Comparison of genetic distance between SNP and SSR markers

580 We calculated the modified Roger's distance (MRD) (Rogers 1972) based on allelic 581 frequency data between landraces using different sets of markers to analyze the effect of the wd 582 criterion (Figure S5) and of the number of markers (Figure S6) on the estimation of relatedness. 583 To analyze the effect of wd criterion, we selected four random sets of 2,000 SNPs with different 584 wd ranges (0-20, 20-40, 40-60, 60-80) among 32,788 PZE SNPs. To analyze the effect of SNP 585 number, we selected six random sets of SNPs with various number of SNPs (15,000, 10,000, 586 5000, 2500, 1000, 500) among 23,412 SNPs with wd above 50. In order to test if the genetic 587 distance is robust when changing the type and the number of markers, we compared MRD 588 between landraces estimated with different SNP datasets with that estimated with 17 SSR 589 markers (Figure 5, Figure S5 and Figure S6). Missing allele frequencies within accession were 590 replaced by corresponding average frequencies within the whole set of accessions before running this analysis. Allelic frequencies of two samples for replicated landraces were averaged 591 592 before estimating MRD distance except for Pol3 for which one of two samples was removed 593 (WG0109808-DNAH04).

594 Coefficient of determination between the distance matrices based on different subsets of 595 SNP (MRD_{SNP}) and 17 SSR markers (MRD_{SSR}) was determined by using linear regression. 596

597 Acknowledgements

598 This study was funded by l'Association pour l'étude et l'amélioration du mais (PROmais) in 599 the project "Diversity Zea" and French National Research Agencies in project Investissement d'Avenir Amaizing, (ANR-10-BTBR-01). We acknowledge greatly the French maize 600 601 Biological Ressources Center, PROmais, and INRAE experimental units of St Martin de Hinx 602 and Mauguio for collecting and maintaining Landraces and Inbred lines collection. We greatly 603 acknowledge the colleagues who initially collected these landraces and André Gallais for 604 having initiated these research programs. We also greatly acknowledge Pierre Dubreuil, Letizia 605 Camus-Kulandaivelu, Cecile Rebourg, Céline Mir, Domenica Maniccaci that conducted 606 previous study on these landraces using DNA pooling approach with SSR and RFLP markers. 607 The Infinium genotyping work was supported by CEA-CNG, by giving the INRAE-EPGV 608 group access to its DNA and cell bank service for DNA quality control and to their Illumina 609 genotyping platform. Thanks respectively to Anne Boland and Marie-Thérèse Bihoreau and 610 their staff. We acknowledge the EPGV group, Dominique Brunel, Marie-Christine Le Paslier, 611 Aurélie Chauveau for the discussion and management of the Illumina genotyping.

612 Author's contribution

- 613 S.D.N, A.C and B.G designed and supervised the study and selected the plant material
- 614 M.A, S.D.N, A.C drafted and corrected the manuscript
- 615 D.M, V.C and A.B extracted DNA and managed genotyping of landraces and inbred lines
- 616 C.B, B.G and A.C collected, maintained landraces, and inbred lines collection
- 617 S.D.N, M.A, A.C and T.M-H developed the statistical methods and scripts for predicting
- 618 allelic frequency from fluorescent data
- 619 M.A, B.G and S.D.N analyzed genetic diversity of landraces panel.
- 620 All authors read and approved the manuscript.

621 Data availability

622 R scripts and fluorescent intensity data of 327 inbred lines and two series of controlled pools 623 used for predicting allelic frequency in DNA bulks of maize landraces by our two-step 624 approaches are available at https://doi.org/10.15454/GANJ7J. Fluorescent Intensity data and 625 allelic frequencies of 20 samples corresponding to 10 duplicated landraces were also available 626 at https://doi.org/10.15454/GANJ7J. Allelic frequencies of new DNA bulks for new maize 627 populations genotyped by maize 50K array could be predicted by using these datasets with R 628 scripts. Note that these datasets and R scripts will become available when the publication would 629 be accepted in a peer review journal.

630 Conflicts of interest

631 No

632

633 **REFERENCES**

- Abe, A., S. Kosugi, K. Yoshida, S. Natsume, H. Takagi *et al.*, 2012 Genome sequencing reveals
 agronomically important loci in rice using MutMap. Nat Biotech 30: 174–178.
- Adoukonou-Sagbadja, H., C. Wagner, A. Dansi, J. Ahlemeyer, O. Daïnou *et al.*, 2007 Genetic
 diversity and population differentiation of traditional fonio millet (Digitaria spp.)
 landraces from different agro-ecological zones of West Africa. Theor. Appl. Genet. 115:
 917–931.
- Aguirre-Liguori, J. A., M. I. Tenaillon, A. Vázquez-Lobo, B. S. Gaut, J. P. Jaramillo-Correa *et al.*, 2017 Connecting genomic patterns of local adaptation and niche suitability in teosintes. Mol. Ecol. 26: 4226–4240.
- Arteaga, M. C., A. Moreno-Letelier, A. Mastretta-Yanes, A. Vázquez-Lobo, A. Breña-Ochoa
 et al., 2016 Genomic variation in recently collected maize landraces from Mexico.
 Genomics Data 7: 38–45.
- Backes, G., B. Hatz, A. Jahoor, and G. Fischbeck, 2003 RFLP diversity within and between
 major groups of barley in Europe. Plant Breed. 122: 291–299.
- Barcellos, L. F., W. Klitz, L. L. Field, R. Tobias, A. M. Bowcock *et al.*, 1997 Association
 Mapping of Disease Loci, by Use of a Pooled DNA Genomic Screen. Am. J. Hum.
 Genet. 61: 734–747.
- Baum, A. E., N. Akula, M. Cabanero, I. Cardona, W. Corona *et al.*, 2007 A genome-wide
 association study implicates diacylglycerol kinase eta (DGKH) and several other genes
 in the etiology of bipolar disorder. Mol. Psychiatry 13: 197–207.
- Bhattacharjee, R., P. Bramel, C. Hash, M. Kolesnikova-Allen, and I. Khairwal, 2002
 Assessment of genetic diversity within and between pearl millet landraces. Theor. Appl.
 Genet. 105: 666–673.
- Bouchet, S., B. Servin, P. Bertin, D. Madur, V. Combes *et al.*, 2013 Adaptation of maize to
 temperate climates: mid-density genome-wide association genetics and diversity
 patterns reveal key genomic regions, with a major contribution of the Vgt2 (ZCN8)
 locus. PloS One 8: e71377.
- Brauner, P. C., W. Schipprack, H. F. Utz, E. Bauer, M. Mayer *et al.*, 2019 Testcross
 performance of doubled haploid lines from European flint maize landraces is promising
 for broadening the genetic base of elite germplasm. Theor. Appl. Genet. 132: 1897–
 1908.
- Breen, G., P. Sham, T. Li, D. Shaw, D. A. Collier *et al.*, 1999 Accuracy and sensitivity of DNA
 pooling with microsatellite repeats using capillary electrophoresis. Mol. Cell. Probes
 13: 359–365.

- Brohede, J., 2005 PPC: an algorithm for accurate estimation of SNP allele frequencies in small
 equimolar pools of DNA using data from high density microarrays. Nucleic Acids Res.
 33: e142–e142.
- 671 Camus-Kulandaivelu, L., 2006 Maize Adaptation to Temperate Climate: Relationship Between
 672 Population Structure and Polymorphism in the Dwarf8 Gene. Genetics 172: 2449–2463.
- 673 Clark, A. G., M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen, 2005
 674 Ascertainment bias in studies of human genome-wide polymorphism. Genome Res. 15:
 675 1496–1502.
- 676 Craig, D. W., M. J. Huentelman, D. Hu-Lince, V. L. Zismann, M. C. Kruer *et al.*, 2005
 677 Identification of disease causing loci using an array-based genotyping approach on
 678 pooled DNA. BMC Genomics 6: 138.
- Daniels, J., P. Holmans, N. Williams, D. Turic, P. McGuffin *et al.*, 1998 A simple method for
 analyzing microsatellite allele image patterns generated from DNA pools and its
 application to allelic association studies. Am. J. Hum. Genet. 62: 1189–1197.
- Dias, J. S., M. B. Lima, K. M. Song, A. A. Monteiro, P. H. Williams *et al.*, 1991 Molecular
 taxonomy of Portuguese tronchuda cabbage and kale landraces using nuclear RFLPs.
 Euphytica 58: 221–229.
- Divaret, I., E. Margalé, and G. Thomas, 1999 RAPD markers on seed bulks efficiently assess
 the genetic diversity of a Brassica oleracea L. collection: Theor. Appl. Genet. 98: 1029–
 1035.
- Dubreuil, P., and A. Charcosset, 1998 Genetic diversity within and among maize populations:
 a comparison between isozyme and nuclear RFLP loci. Theor. Appl. Genet. 96: 577–
 587.
- Dubreuil, P., C. Rebourg, M. Merlino, and A. Charcosset, 1999 Evaluation of a DNA pooled sampling strategy for estimating the RFLP diversity of maize populations. Plant Mol
 Biol Rep 17: 123–138.
- Dubreuil, P., M. Warburton, M. Chastanet, D. Hoisington, and A. Charcosset, 2006 More on
 the introduction of temperate maize into Europe: Large-scale bulk SSR genotyping and
 new historical elements. Maydica 51: 281–291.
- Elferink, M. G., H.-J. Megens, A. Vereijken, X. Hu, R. P. M. A. Crooijmans *et al.*, 2012
 Signatures of Selection in the Genomes of Commercial and Non-Commercial Chicken
 Breeds (T. Shioda, Ed.). PLoS ONE 7: e32720.
- Eschholz, T. W., P. Stamp, R. Peter, J. Leipner, and A. Hund, 2010 Genetic structure and history
 of Swiss maize (Zea mays L. ssp. mays) landraces. Genet. Resour. Crop Evol. 57: 71–
 84.
- Fernie, A. R., Y. Tadmor, and D. Zamir, 2006 Natural genetic variation for improving crop quality. Curr. Opin. Plant Biol. 9: 196–202.

- Ford-Lloyd, B. V., H. J. Newbury, M. T. Jackson, and P. S. Virk, 2001 Genetic basis for co adaptive gene complexes in rice (Oryza sativa L.) landraces. Heredity 87: 530–536.
- Frascaroli, E., T. A. Schrag, and A. E. Melchinger, 2013 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. Theor. Appl. Genet. 126: 133–141.
- Gallais, A., L. Moreau, and A. Charcosset, 2007 Detection of marker–QTL associations by
 studying change in marker frequencies with selection. Theor. Appl. Genet. 114: 669–
 681.
- Ganal, M. W., G. Durstewitz, A. Polley, A. Bérard, E. S. Buckler *et al.*, 2011 A Large Maize
 (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and
 Genetic Mapping to Compare with the B73 Reference Genome (L. Lukens, Ed.). PLoS
 ONE 6: e28334.
- Gauthier, P., B. Gouesnard, J. Dallard, R. Redaelli, C. Rebourg *et al.*, 2002 RFLP diversity and
 relationships among traditional European maize populations. Theor Appl Genet 105:
- Gautier, M., J. Foucaud, K. Gharbi, T. Cézard, M. Galan *et al.*, 2013 Estimation of population
 allele frequencies from next-generation sequencing data: pool-versus individual-based
 genotyping. Mol. Ecol. 22: 3766–3779.
- Glaubitz, J. C., T. M. Casstevens, F. Lu, J. Harriman, R. J. Elshire *et al.*, 2014 TASSEL-GBS:
 A High Capacity Genotyping by Sequencing Analysis Pipeline (N. A. Tinker, Ed.).
 PLoS ONE 9: e90346.
- Hagenblad, J., J. Zie, and M. W. Leino, 2012 Exploring the population genetics of genebank
 and historical landrace varieties. Genet. Resour. Crop Evol. 59: 1185–1199.
- Hamblin, M. T., M. L. Warburton, and E. S. Buckler, 2007 Empirical Comparison of Simple
 Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize
 Diversity and Relatedness (H. Ellegren, Ed.). PLoS ONE 2: e1367.
- van Heerwaarden, J., J. Doebley, W. H. Briggs, J. C. Glaubitz, M. M. Goodman *et al.*, 2011
 Genetic signals of origin, spread, and introgression in a large sample of maize landraces.
 Proc. Natl. Acad. Sci. 108: 1088–1092.
- Hoisington, D., M. Khairallah, T. Reeves, J.-M. Ribaut, B. Skovmand *et al.*, 1999 Plant genetic
 resources: What can they contribute toward increased crop productivity? Proc. Natl.
 Acad. Sci. 96: 5937–5943.
- Hölker, A. C., M. Mayer, T. Presterl, T. Bolduan, E. Bauer *et al.*, 2019 European maize
 landraces made accessible for plant breeding and genome-based studies. Theor. Appl.
 Genet. 132: 3333–3345.
- Hoogendoorn, B., N. Norton, G. Kirov, N. Williams, M. Hamshere *et al.*, 2000 Cheap, accurate
 and rapid allele frequency estimation of single nucleotide polymorphisms by primer
 extension and DHPLC in DNA pools. Hum. Genet. 107: 488–493.

- Inghelandt, D., J. C. Reif, B. S. Dhillon, P. Flament, and A. E. Melchinger, 2011 Extent and
 genome-wide distribution of linkage disequilibrium in commercial maize germplasm.
 Theor. Appl. Genet. 123: 11–20.
- Jawaid, A., and P. Sham, 2009 Impact and Quantification of the Sources of Error in DNA
 Pooling Designs. Ann. Hum. Genet. 73: 118–124.
- Johnston, S. E., M. Lindqvist, E. Niemelä, P. Orell, J. Erkinaro *et al.*, 2013 Fish scales and SNP
 chips: SNP genotyping and allele frequency estimation in individual and pooled DNA
 from historical samples of Atlantic salmon (Salmo salar). BMC Genomics 14: 439.
- Kilian, B., and A. Graner, 2012 NGS technologies for analyzing germplasm diversity in genebanks*. Brief. Funct. Genomics 11: 38–50.
- Kirov, G., I. Nikolov, L. Georgieva, V. Moskvina, M. J. Owen *et al.*, 2006 Pooled DNA
 genotyping on Affymetrix SNP genotyping arrays. BMC Genomics 7: 27.
- Lam, H.-M., X. Xu, X. Liu, W. Chen, G. Yang *et al.*, 2010 Resequencing of 31 wild and
 cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat.
 Genet. 42: 1053–1059.
- Laval, G., M. SanCristobal, and C. Chevalet, 2002 Measuring genetic distances between breeds:
 use of some distances in various short term evolution models. Genet. Sel. Evol. 34: 481–
 507.
- Le Hellard, S., S. J. Ballereau, P. M. Visscher, H. S. Torrance, J. Pinson *et al.*, 2002 SNP
 genotyping on pooled DNAs: comparison of genotyping technologies and a semi
 automated method for data storage and analysis. Nucleic Acids Res. 30: e74–e74.
- LeDuc, C., P. Miller, J. Lichter, and P. Parry, 1995 Batched analysis of genotypes. Genome
 Res. 4: 331–336.
- Lipkin, E., M. O. Mosig, A. Darvasi, E. Ezra, A. Shalom *et al.*, 1998 Quantitative trait locus
 mapping in dairy cattle by means of selective milk DNA pooling using dinucleotide
 microsatellite markers: analysis of milk protein percentage. Genetics 149: 1557–1567.
- Liu, K., M. Goodman, S. Muse, J. S. Smith, E. Buckler *et al.*, 2003 Genetic structure and
 diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:
 2117–2117.
- Mascher, M., M. Schreiber, U. Scholz, A. Graner, J. C. Reif *et al.*, 2019 Genebank genomics
 bridges the gap between the conservation of crop diversity and plant breeding. Nat.
 Genet. 51: 1076–1081.
- McCouch, S. R., K. L. McNally, W. Wang, and R. Sackville Hamilton, 2012 Genomics of gene banks: A case study in rice. Am. J. Bot. 99: 407–423.
- McCouch, S. R., K. Zhao, M. Wright, C.-W. Tung, K. Ebana *et al.*, 2010 Development of
 genome-wide SNP assays for rice. Breed. Sci. 60: 524–535.

- Mir, C., T. Zerjal, V. Combes, F. Dumas, D. Madur *et al.*, 2013 Out of America: tracing the genetic footprints of the global diffusion of maize. Theor. Appl. Genet. 126: 2671–2682.
- Moragues, M., J. Comadran, R. Waugh, I. Milne, A. J. Flavell *et al.*, 2010 Effects of
 ascertainment bias and marker number on estimations of barley diversity from highthroughput SNP genotype data. Theor. Appl. Genet. 120: 1525–1534.
- Nielsen, R., 2004 Population genetic analysis of ascertained SNP data. Hum Genomics 1: 218–
 224.
- Ozerov, M., A. Vasemägi, V. Wennevik, R. Diaz-Fernandez, M. Kent *et al.*, 2013 Finding
 Markers That Make a Difference: DNA Pooling and SNP-Arrays Identify Population
 Informative Markers for Genetic Stock Identification (S. Consuegra, Ed.). PLoS ONE
 8: e82434.
- Parzies, H. K., W. Spoor, and R. A. Ennos, 2000 Genetic diversity of barley landrace accessions
 (Hordeum vulgare ssp. vulgare) conserved for different lengths of time in ex situ gene
 banks. Heredity 84: 476–486.
- Peiris, B. L., J. Ralph, S. J. Lamont, and J. C. M. Dekkers, 2011 Predicting allele frequencies
 in DNA pools using high density SNP genotyping data. Anim. Genet. 42: 113–116.
- Perlin, M. W., G. Lancia, and S.-K. Ng, 1995 Toward fully automated genotyping: genotyping
 microsatellite markers by deconvolution. Am. J. Hum. Genet. 57: 1199.
- Pervaiz, Z. H., M. A. Rabbani, I. Khaliq, S. R. Pearce, and S. A. Malik, 2010 Genetic diversity
 associated with agronomic traits using microsatellite markers in Pakistani rice
 landraces. Electron J Biotechnol 13:.
- Pupilli, F., P. Labombarda, C. Scotti, and S. Arcioni, 2000 RFLP analysis allows for the
 identification of alfalfa ecotypes. Plant Breed. 119: 271–276.
- Rafalski, A., 2002 Applications of single nucleotide polymorphisms in crop genetics. Curr Opin
 Plant Biol 5: 94–100.
- Rebourg, C., M. Chastanet, B. Gouesnard, C. Welcker, P. Dubreuil *et al.*, 2003 Maize
 introduction into Europe: the history reviewed in the light of molecular data. Theor Appl
 Genet 106: 895–903.
- Rebourg, C., P. Dubreuil, and A. Charcosset, 1999 Genetic diversity among maize populations:
 bulk RFLP analysis of 65 accessions. Maydica 44: 237–249.
- Rebourg, C., B. Gouesnard, and A. Charcosset, 2001 Large scale molecular analysis of traditional European maize populations. Relationships with morphological variation.
 Heredity 86: 574–587.
- Reif, J. C., S. Hamrit, M. Heckenberger, W. Schipprack, H. Peter Maurer *et al.*, 2005a Genetic
 structure and diversity of European flint maize populations determined with SSR
 analyses of individuals and bulks. Theor. Appl. Genet. 111: 906–913.

- Reif, J. C., M. L. Warburton, X. C. Xia, D. A. Hoisington, J. Crossa *et al.*, 2006 Grouping of
 accessions of Mexican races of maize revisited with SSR markers. Theor. Appl. Genet.
 113: 177–185.
- Reif, J. C., P. Zhang, S. Dreisigacker, M. L. Warburton, M. van Ginkel *et al.*, 2005b Wheat
 genetic diversity trends during domestication and breeding. Theor. Appl. Genet. 110:
 819 859–864.
- Reyes-Valdés, M. H., A. Santacruz-Varela, O. Martínez, J. Simpson, C. Hayano-Kanashiro *et al.*, 2013 Analysis and Optimization of Bulk DNA Sampling with Binary Scoring for
 Germplasm Characterization (T. Zhang, Ed.). PLoS ONE 8: e79936.
- Rincent, R., L. Moreau, H. Monod, E. Kuhn, A. E. Melchinger *et al.*, 2014 Recovering Power
 in Association Mapping Panels with Variable Levels of Linkage Disequilibrium.
 Genetics 197: 375–387.
- Rogers, J. S., 1972 Measures of genetic similarity and genetic distance. Stud. Genet. 7: 145–
 153.
- Rousselle, Y., E. Jones, A. Charcosset, P. Moreau, K. Robbins *et al.*, 2015 Study on Essential
 Derivation in Maize: III. Selection and Evaluation of a Panel of Single Nucleotide
 Polymorphism Loci for Use in European and North American Germplasm. Crop Sci.
 55: 1170.
- Schlötterer, C., R. Tobler, R. Kofler, and V. Nolte, 2014 Sequencing pools of individuals —
 mining genome-wide polymorphism data without big funding. Nat. Rev. Genet. 15:
 749–763.
- 835 Segovia-Lerma, A., R. G. Cantrell, J. M. Conway, and I. M. Ray, 2003 AFLP-based assessment
 836 of genetic diversity among nine alfalfa germplasms using bulk DNA templates. Genome
 837 46: 51–58.
- Sham, P., J. S. Bader, I. Craig, M. O'Donovan, and M. Owen, 2002 DNA Pooling: a tool for
 large-scale association studies. Nat. Rev. Genet. 3: 862–871.
- Simpson, C. L., 2005 A central resource for accurate allele frequency estimation from pooled
 DNA genotyped on DNA microarrays. Nucleic Acids Res. 33: e25–e25.
- Strigens, A., W. Schipprack, J. C. Reif, and A. E. Melchinger, 2013 Unlocking the Genetic
 Diversity of Maize Landraces with Doubled Haploids Opens New Avenues for
 Breeding (R. P. Niedz, Ed.). PLoS ONE 8: e57234.
- Tanksley, S. D., 1997 Seed Banks and Molecular Maps: Unlocking Genetic Potential from the
 Wild. Science 277: 1063–1066.
- Teumer, A., F. D. Ernst, A. Wiechert, K. Uhr, M. Nauck *et al.*, 2013 Comparison of genotyping
 using pooled DNA samples (allelotyping) and individual genotyping using the
 affymetrix genome-wide human SNP array 6.0. BMC Genomics 14: 506.

- Vigouroux, Y., S. Mitchell, Y. Matsuoka, M. Hamblin, S. Kresovich *et al.*, 2005 An analysis
 of genetic diversity across the maize genome using microsatellites. Genetics 169: 1617–
 1630.
- Yang, X., Y. Xu, T. Shah, H. Li, Z. Han *et al.*, 2011 Comparison of SSRs and SNPs in assessment of genetic relatedness in maize. Genetica 139: 1045–1054.
- Yao, Q., K. Yang, G. Pan, and T. Rong, 2007 Genetic diversity of maize (Zea mays L.)
 landraces from Southwest China based on SSR data. J. Genet. Genomics 34: 851–860.
- Zhao, W., 2006 Panzea: a database and resource for molecular and functional diversity in the
 maize genome. Nucleic Acids Res. 34: D752–D757.

859

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103655. this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 1: Expected frequencies of allele B for the nine controlled pools obtained by varying the proportions of leaf weights of three inbred lines (L1, L2, L3) according to their genotypes at a bi-allelic SNP coded A/B. Heterozygous genotypes for inbred lines were not considered in this table.

Pools	Propor w	tion of eights	leaf	Genotypes of parental lines L1, L2, L3 in controlled pools							
		L1			<u>.</u>	AA,AA,BB	BB,AA,BB	AA,BB,BB	BB,BB,BB		
	L1		L3	AA,AA,AA	BB,AA,AA	OR	or				
						AA,BB,AA	BB,BB,AA				
#1	0.01	0.495	0.495	0%	1%	50%	51%	99%	100%		
#2	0.02	0.49	0.49	0%	2%	49%	51%	98%	100%		
#3	0.03	0.485	0.485	0%	3%	49%	52%	97%	100%		
#4	0.05	0.475	0.475	0%	5%	48%	53%	95%	100%		
#5	0.07	0.465	0.465	0%	7%	47%	54%	93%	100%		
#6	0.1	0.45	0.45	0%	10%	45%	55%	90%	100%		
#7	0.15	0.425	0.425	0%	15%	43%	58%	85%	100%		
#8	0.2	0.4	0.4	0%	20%	40%	60%	80%	100%		
#9	0.333	0.333	0.333	0%	33%	33%	67%	67%	100%		
Configuration of controlled pools			olled	Monomorphic	R1	R2	R3		Monomorphic		

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103655. this version posted May 19, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.

Table 2: Mean absolute error (MAE) in frequency estimation for 1,000 SNPs used to calibrate logistic regressionequations. MAE is estimated by a cross-validation procedure in which a number of pools comprised between 1 and 15among 18 is removed at random from the calibration set. This procedure was repeated 1,000 times for each SNP.

# of	removed	#	of	Mean absolute				
samples		repetitions	-	Mean	$\frac{IAE}{SD^*}$			
	1	1000		0.0414	0.0219			
	3	1000		0.0428	0.0226			
	5	1000		0.0447	0.0232			
	8	1000		0.0484	0.0245			
	10	1000		0.0522	0.0257			
	12	1000		0.0582	0.0274			
	15	1000		0.0854	0.0309			

* SD = Standard deviation

	15 individuals							100 individuals						
	One biological			Tw	Two biological		One biological			Tw	Two biological			
	replicate			replicates			replicate			replicates				
Allelic	elic # Lower Upper		#	Lower	Upper	# Lower Upper		#	Lower	Upper				
Frequency	Alleles	bound	bound	Alleles	bound	bound	Alleles	bound	bound	Alleles	bound	bound		
0	0	0	0.116	0	0	0.06	0	0	0.018	0	0	0.009		
0.03	1	0.001	0.172	2	0.004	0.115	6	0.011	0.064	13	0.017	0.055		
0.1	3	0.021	0.265	6	0.038	0.205	20	0.062	0.15	40	0.072	0.134		
0.2	6	0.077	0.386	12	0.108	0.323	40	0.147	0.262	80	0.162	0.243		
0.3	9	0.147	0.494	18	0.189	0.432	60	0.237	0.369	120	0.256	0.348		
0.4	12	0.227	0.594	24	0.276	0.535	80	0.332	0.472	160	0.352	0.45		
0.5	15	0.313	0.687	30	0.368	0.632	100	0.429	0.571	200	0.45	0.55		
0.6	18	0.406	0.773	36	0.465	0.724	120	0.529	0.669	240	0.55	0.648		
0.7	21	0.506	0.853	42	0.568	0.812	140	0.631	0.763	280	0.653	0.745		
0.8	24	0.614	0.923	48	0.677	0.892	160	0.738	0.853	320	0.757	0.838		
0.9	27	0.735	0.979	54	0.795	0.962	180	0.85	0.938	360	0.866	0.928		
1	30	0.884	1	60	0.94	1	200	0.982	1	400	0.991	1		

Table 3: Sampling error estimated by numerical calculation for one or two biological replicates with independent sampling of 15 or 100 individuals within landraces. Lower and upper bounds indicate the 95% confidence interval for the allelic frequency in the population, based on the binomial probability of the frequency estimated with the corresponding sample size.



Figure 1: Two-step approach for estimating allelic frequency in DNA pools, exemplified by marker PZE-101005765. Red and green histograms correspond to the fluorescent intensity ratio (FIR) distribution for inbred lines homozygote for allele A (AA) and B (BB), respectively. Red and green curves indicate the corresponding Gaussian distributions. Red, Blue, and Green areas correspond to the FIR for which landraces are declared homozygous for allele A, polymorphic and homozygous for allele B after testing for fixation of alleles A and B. Dotted blue line corresponds to the curve of the logistic regression adjusted on 1,000 SNPs and two series of controlled pools. Blue cross corresponds to a landrace represented by a DNA bulk of 15 individuals, with its observed FIR on X axis and predicted frequency on Y axis.



Figure 2: Mean absolute error (MAE) according to the known allelic frequency in two series of controlled pools. MAE measured the absolute difference between allelic frequencies predicted by the two-step approach and those expected from the genotypes of parental lines in two series of controlled pools for 23,412 SNPs. MAE is averaged for each interval of expected allelic frequency across all SNPs.



Figure 3: Relationship between fluorescent intensity ratio of European Flint controlled pools genotyped in two different laboratories: CNG and Consortium. Each dot represents the combination of one out 9 controlled pools and one out of 23,412 PZE SNPs. Coefficient of determination (r²) between FIR of two laboratories is 0.987.



Figure 4: Relationship between allele frequencies predicted for two biological replicates of 9 landraces over 23,412 selected SNPs. Each dot represents one landrace and one SNP, with allele frequency of replicates 1 and 2 on X and Y axes, respectively. Blue line indicates linear regression. 94.5% of points are included in the red ellipse that represents the 95% confidence limit accounting for the effect of sampling alone. r² between replicates is 0.93



Figure 5: Relationship between Modified Roger's Distances (MRD) obtained with 17 SSRs and 23,412 SNPs for 156 landraces. Each dot represents one pair of landraces. Red dotted lines represents linear regression between MRD_{SSR} and MRD_{SNP}. Coefficient of determination (r²) is reported on the plot.