Seeds of Discovery (**SeeD**): Next Generation Applications to Explore Maize Genetic Diversity

MasAgro
Modernización Sustentable de la Agricultura Tradicional

Sarah J. Hearne[1*], Juan Burgueño[1], J. Andrés Christen[2], Gilberto Esquivel[3], Jorge Franco[4], Juan Manuel Hernández Casillas[3], Andrzej Kilian[6], Ky L. Mathews[1], Peter Wenzl, Martha C. Willcox[1] *1 CIMMYT, 2 CIMAT, 3 INIFAP, 4 Universidad de la Republica Uruguay, 5 Diversity Arrays Technology Pty Ltd (DArT PL). * shearne@cgiar.org*

Introduction

The collaborative Seeds of Discovery initiative (SeeD) aims to enable the study and utilization of maize landrace and wild relative accessions present in the CIMMYT international maize germplasm bank and partnering national germplasm collections within Mexico. One objective within the initiative is to create a global molecular atlas of maize.

Using genotyping by sequencing (GBS) we aim to generate genetic profiles of more than 40,000 maize landraces, advanced breeding lines and wild relatives. Maize landrace accessions, in contrast to breeding materials, are typically heterozygous and highly heterogeneous in nature. In order to accurately represent the fingerprint of an accession using existing GBS approaches it is necessary to genotype multiple individuals using data generation methods which enable the direct scoring of heterozygous states (achieved through reduced multiplexing or increasing the number of sequencing runs of any one sample). Imputation in landrace materials, unlike breeding germplasm offers more challenges as pedigree structure is unknown, though IBD could be used as a proxy. The direct scoring of heterozygotes is a desirable feature at this time of any GBS approach. The need for multiple individuals and deep sequencing significantly increases the cost and time to conduct landrace fingerprinting using current GBS approaches. Here we describe the development of a modified GBS application optimized to provide cost effective, scientifically relevant, high throughput genotyping of maize landraces.

Methods and Results

The choice of enzyme(s) used to prepare a reduced complexity library for NGS is critical. To enable accurate scoring of heterozygotes a suitable depth of sequencing is required and this can be achieved through reduction in the number of fragments generated during library preparation.
Briefly ten different library complexity reduction methods (restriction enzyme combinations) were evaluated to investigate applicability to maize. Of these three were further evaluated to determine the number of fragments generated, the sequencing depth across loci and the repeatability of the genotyping across DNA samples, DNA barcodes and sequencing runs. Twenty four individuals from each of eight diverse accessions from the INIFAP genebank were used as samples. DNA was extracted using the CTAB method from lyophilized leaves harvested from young field grown plants. The same DNA samples were used for each complexity reduction method. Of the three methods evaluated the second method provided the highest number of 10x coverage markers (Table 1) and in additionally produced high repeatability across DNA sample, barcode and sequencing run (Figure 1)

Method	Mt and plastid loci	Number of loci >10X	Number of loci >20X	Number of loci >30X	Number of loci >40X
1	3% of total counts	8860	2150	540	200
2	0.02% of total counts	9190	3940	1786	700
3	0.01% of total counts	2460	1930	1640	1430

Table 1. The number of SNP sequenced at 10, 20, 30 and 40X coverage using three different complexity reduction methods. The proportion of mitochondrial and plastid fragments sequenced as a proportion of the total number of counts is presented

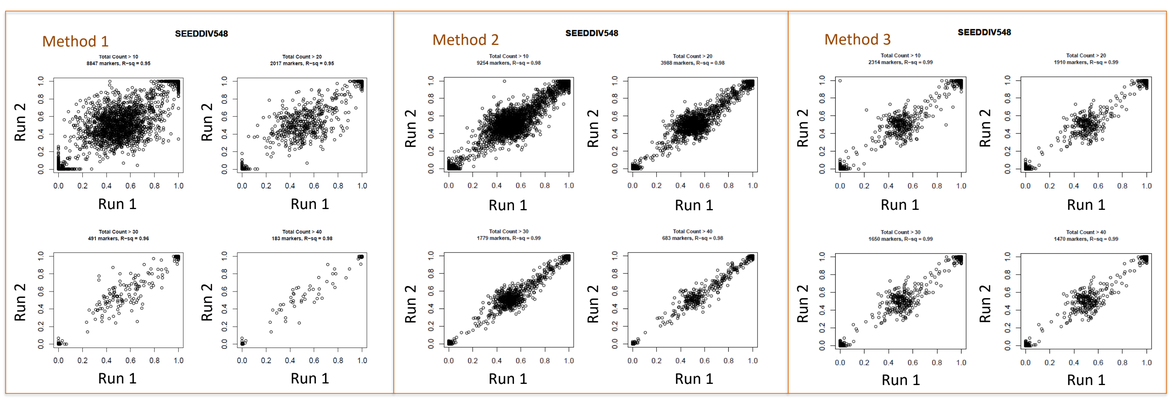


Figure 1. Repeatability of GBS SNP allele frequency determined from number of counts across two sequencing runs using three different methods of complexity reduction. Data presented are split into four coverage categories.

The ability to combine multiple individuals into one sample, while being able to accurately estimate allele frequency would enable the resource efficient representation of landrace-level fingerprints and enable resources to be spent on expanding the range of accessions evaluated compared with the scenario of evaluating multiple individuals. In order to evaluate the effectiveness of the three methods to estimate allele frequencies in combined samples DNA bulks (single extractions from leaf of multiple individuals) and pools (equimolar pools of DNA from individuals) from the four INIFAP accessions were generated with sample sizes ranging from 4 individuals too 48 individuals (DNA samples with >24 individuals were made from two accessions). GBS was conducted using all methods and the proportion of the reference allele (based on count) in the pools and bulks was compared with the mean proportion of the allele present in the corresponding individuals . Figure 2 indicates the divergence of allele frequencies estimated from pools or bulks and corresponding individuals for all three complexity reduction methods. As indicated the divergence seen in method two was significantly lower than that seen in the other two approaches.

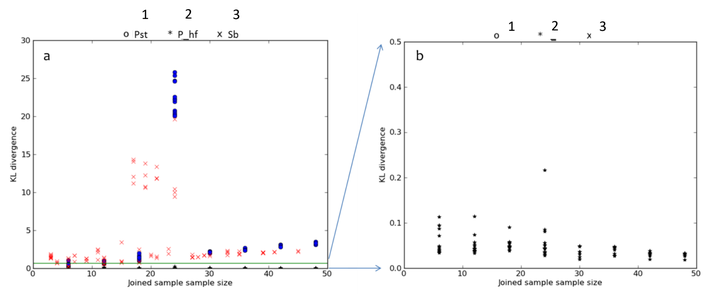
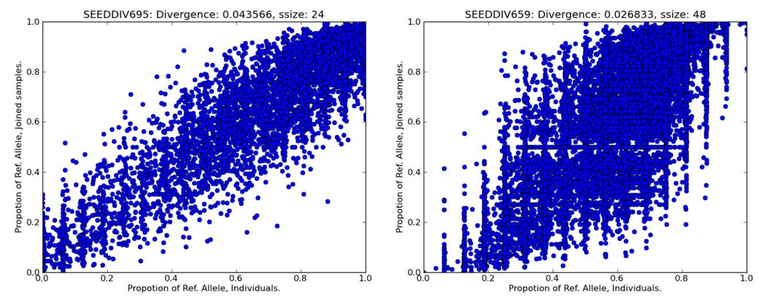


Figure 2. Relationship between the proportion of reference allele in pooled and bulk DNA samples (joined samples) compared with the proportion of reference allele in corresponding individual samples. Data represent individual SNP sequenced at >10X coverage.

Method two was investigated further to better evaluate the representativeness of a pooled or bulk methods to estimate allele frequencies across accessions. As figure 3 indicates the bulk and pool methods provided good representation of the allele frequencies found in individuals across all sample sizes evaluated. There were no significant influence of DNA preparation method and due to experimental simplicity the bulk method was chosen for further evaluations of GBS of multiple individuals.

Figure 3. Relationship between the proportion of reference allele in pooled and bulk DNA samples (joined samples) compared with the proportion of reference allele in corresponding individual samples. Data represent individual SNP sequenced at >10X coverage.



The optimal number of individuals to sample per accession was evaluated via GBS (method two) of a minimum of 92 individuals from each of 8 diverse accessions obtained from the CIMMYT genebank. This was complemented by GBS of bulks from each accessions ranging from 4 to 48 individuals forming two independent bulks per accession. Divergence between independent bulks stabilized around sample sizes of 20 individuals (Figure 4). The measure of genetic distance was largely representative of the population mean at low number of individuals (data not shown), however diversity indices were greatly influenced by sample size. Linear and quadratic segmental regression with plateau was used on He and proportion of polymorphic loci measures to identify optimal numbers of individuals per accessions. Sample sizes ranging from 12 to 33 provided adequate estimates of diversity in six accessions. In two plateau was not achieved across the sample sizes evaluated (data not shown).

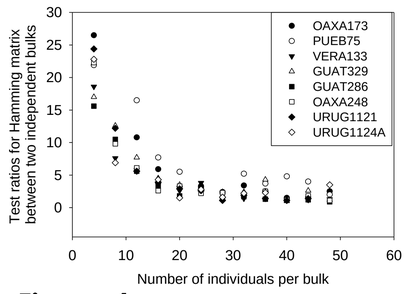


Figure 4. Test ratios for Hamming matrix between independent bulks across eight accessions and bulk sizes from 4 to 48

Conclusions and future work

One complexity reduction method has been identified which produces repeatable fingerprints with high levels of heterozygote calling. Further this approach is applicable to the evaluation of DNA samples derived from multiple individuals providing good resolution of allele frequencies in the sample relative to that of component individuals. Evaluation of an optimal number of individuals to sample per accession indicated that few samples could be used to estimate genetic distance with more required to obtain accessions relevant diversity measures. Thirty plants per accessions provides a resource and information optimized number of individuals to use in DNA bulks, with the ability to use one bulk per accession to obtain a representative landrace genotypic profile. Work has progressed to use this method to conduct GBS on 2000 accessions to date and a further 12000 are in the DNA extraction pipeline (Figure 5). Work is continuing to understand, partition and document the error incurred using this sampling approach through continued analysis of existing data and the analysis of DH maize lines and F1 progeny of diallels. Data generated through this analysis will be made available through the SeeD project web portal via the molecular atlas of maize.

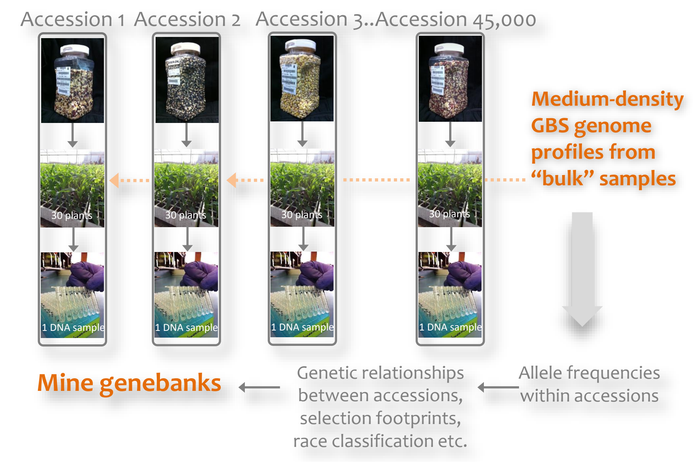


Figure 5. Schema of SeeD-maize genebank accession genotyping

This work is funded through the MASAGRO initiative of SAGARPA and Strategic Initiative 8 of the maize CGIAR research program.

CIMMYT MR
International Maize and Wheat Improvement Center

SAGARPA
SECRETARÍA DE AGRICULTURA, GANADERÍA, DESARROLLO RURAL, PESCA Y ALIMENTACIÓN