Genomic Exploration of Genebank Collections: Early Insights from Seeds of Discovery-Maize

Sarah Hearne[1]*, Jorge Franco[2], Jiafa Chen[1], Cesar Petroli[1]. *1 CIMMYT, 2 Universidad de la Republica Uruguay. * shearne@cgiar.org*

Introduction

At a time of unprecedented challenges to crop production the need to explore and use all available improvement mechanisms is paramount. One intervention available to scientists is the use of native genetic variation. For the major staple commodities the broadest reserves of this variation are typically housed in germplasm banks. There are a number of real and perceived barriers to the use of germplasm from bank collections. One barrier is the limited availability of data pertaining to the materials held in the bank; characteristics be they general or for a specific phenotype. Here we present work from the Seeds of Discovery project (SeeD), focusing on the comprehensive genotypic exploration and targeted use of specific genetic variation from the world's broadest internationally available collection of maize germplasm.

Methods and Results

Genotypic characterisation

All maize accessions within the International Genebank of CIMMYT have been genotyped using a composite genotyping by sequencing (GbS) approach (DArTSeq) specifically developed in collaboration with Diversity Arrays Technology for the assessment of heterogeneous populations along with CIMMYT donor lines and a series of Ex-PVP lines.

- DNA from a composite of 30 individuals per accession / sample was used for GbS employing Illumina HiSeq platforms.
- Resulting sequence reads are clustered to an evolving internal reference of clusters.
- SNP identified using a call rate of 0.01 and the number of sequence read variants per SNP are determined.
- BLAST is conducted anchoring SNP to the most recent build of B73
- SNP "count" data is then converted into a frequency of each allele, per locus, per sample.
- Statistical analysis was conducted in R with multidimensional scaling results visualised in CurlyWhirly.

Analysis presented here is based on 27712 samples which are separated into germplasm types; landraces, pools, populations and varieties, lines, teosintes and *Tripsacum sp.*, lines, teosinte and *Tripsacum sp.* are further sub divided to reflect source or species.

Broad analysis

Across the germplasm evaluated 616967 SNP were identified. The mean number of SNP detected per sample was 258213 with a median of 265940. Approximately half of the SNP can be aligned to B73 (51%), with 44% of the SNP having single alignments. The mean coverage for SNP was 6.85 with a minimum of 0.75 and a maximum of 195.3. Across the different germplasm types mean coverage varied (Table 1). Without marker filtering mean scaled observed heterozygosity (He) across all samples was 0.14 with variation observed across different germplasm types, lines having the lowest observed heterozygosity as expected. As SNP are filtered to remove lower coverage markers and markers with low MAF He increases as anticipated (data not shown).

Table 1. Summary of observed heterozygosity, number of effective genotypes and coverage for each germplasm group analyzed.

Germplasm Type	Number of entries	Mean He	Mean nefgen	Mean Coverage
Landrace pool pop	25833	0.139	10878	5.08
Lines	1498	0.088	551	9.92
Teosinte	231	0.123	128	6.19
Tripsacum	150	0.149	61	8.22

Multidimensional scaling of genetic distance (Fig. 1), indicates a clear separation of *Tripsacum sp.* from improved materials from the landraces present in the genebank along the first dimension (Fig1a). The maize landraces, pool and population group covers the broadest swathe of genetic space across the second and third dimensions (y,z) with the lines and teosinte accessions present forming a discrete interface on the first and third dimensions (Fig1,a,b).

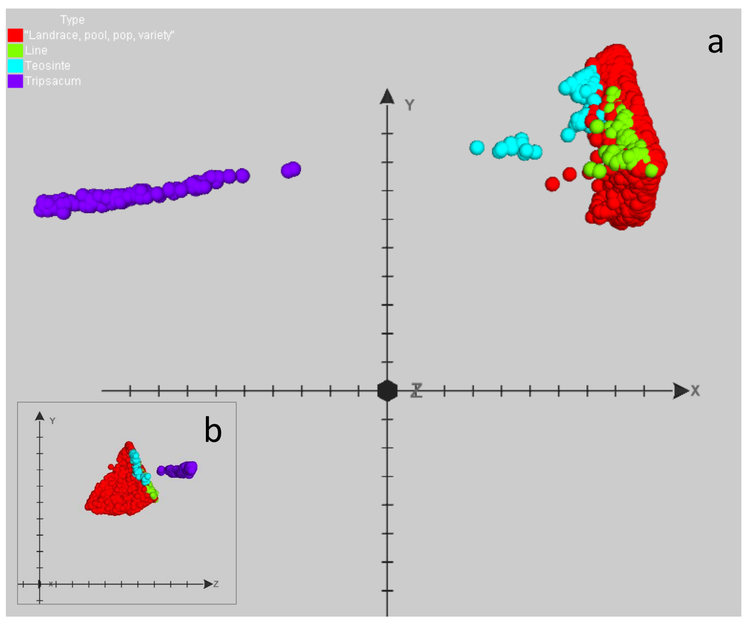


Figure 1. Multidimensional scaling of Modified Rogers distance between 27712 accession and line samples. Points are colored to represent germplasm type a) x,y,z orientation of axes, b) orientation to show distribution along z,y,x axes

The line group of germplasm comprises the elite CIMMYT maize lines (CML) and CIMMYT trait donor lines which have adaptation to tropical lowland, mid-altitude and highland environments. In addition, a group of Ex-PVP temperate lines was included to facilitate direct comparison. The multidimensional scaling (Fig 2) indicates that among the lines the CIMMYT derived materials occupies a larger area of genetic space compared with the ex-PVP materials. The ex-PVP germplasm forms a band along one side of the first dimension (Fig2), with discrete grouping towards the centre of the third dimension(data not shown).This is not unexpected given the broader of agroecologies covered as targets for the CIMMYT maize breeding program and the more diverse range of founder germplasm reflected in these entries. The CIMMYT donor lines are interspersed among the CMLs reflecting the similar and shared nature of founding populations.

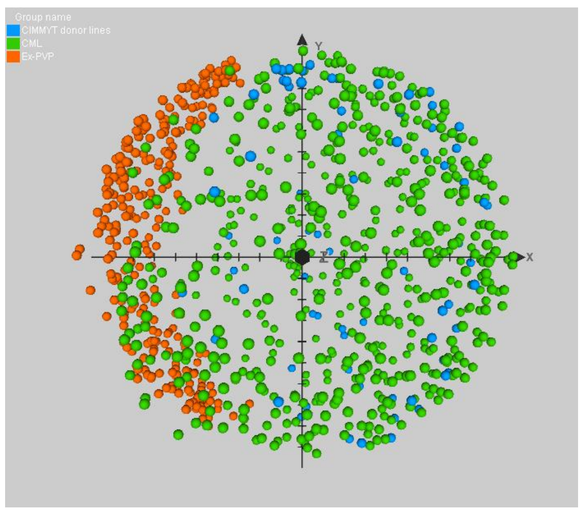


Figure 2. Multidimensional scaling of Modified Rogers distance between 1498 maize lines . Points are colored to represent line type.

Teosinte germplasm separates in genetic space with *Zea luxurians* from Nicaragua grouping clearly from the Mexican *Zea mexicana* accessions (Fig 3). The two Zea diploperennis accessions that have been co-analysed to date locate with *Zea luxurians* on the first dimension but clearly separate on the Z dimension. Using the current grouping it is possible to allocate putative species identity to the unknown samples. This will be verified in the field in due course.

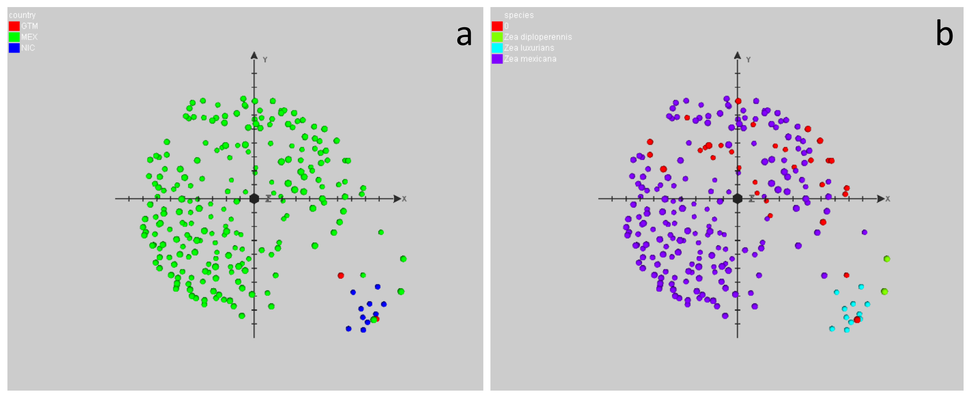
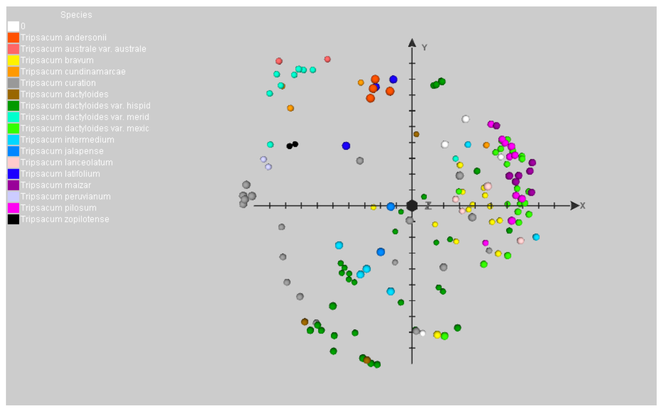


Figure 3. Multidimensional scaling of Modified Rogers distance between 231 teosinte accessions. Points are colored to represent a) country of origin and, b) species.

Preliminary analysis of the wide number of *Tripsacum* species represented in the CIMMYT Genebank illustrates some discrete grouping in some species and broader genetic distance in others (Fig 4). The results need to be interpreted cautiously to reflect the small sample sizes of some species.

Figure 4. Multidimensional scaling of Modified Rogers distance between 150 *Tripsacum sp*. accessions. Points are colored to represent species.



Summary

Analysis presented here demonstrates the clear separation in genetic space of *Zea* from *Tripsacum* species reflecting the ancient speciation event. The wide genetic distance and high diversity present in the *Tripsacum sp*. is apparent reflecting the wide diversity in this genus. A second less dramatic definition of teosintes from *Zea mays* is also clearly evident. Comparison of line germplasm with landrace and population founders demonstrates the impact of genetic selection in limiting the overall genetic distance represented by the lines. Together with GIS and phenotypic data this provides clear evidence of the potential opportunities for sourcing novel alleles of breeding value from genebank source.

Acknowledgements

This work was conducted under the Seeds of Discovery Project supported by SAGARPA (La Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y Alimentación), Mexico under the MasAgro (Sustainable Modernization of Traditional Agriculture) initiative.