ARTÍCULO DE INVESTIGACIÓN - EDICIÓN ESPECIAL

Vol. 40 | No. 1 | **ENERO - ABRIL 2019** | pp 1-17

REVISTA MEXICANA DE Ingeniería Biomédica

dx.doi.org/10.17488/RMIB.40.1.7

E-LOCATION ID: **e201807EE1**

Signal-Processing tools for core-collection selection from genetic-resource collection

Herramienta de procesamiento de señales para la selección de "Core-Collection" desde colección de recursos genéticos

L. I. López-Flores[1], *Masaru Takeya*[2], *E. Borrayo*[1]

[1]Universidad de Guadalajara

[2]National Agriculture and Food Research Organization

ABSTRACT

Selecting a representative core collection (CC) is a proven and effective strategy for overcoming the expenses and difficulties of managing genetic resources in gene banks around the globe. Because of the diverse applications available for these sub-collections, several algorithms have been successfully implemented to construct them based on genotypic, phenotypic, passport or geographic data (either by individual datasets or by consensus). However, to the best of our knowledge, no single comprehensive datasets has been properly explored to date. Thus, researchers evaluate multiple datasets in order to construct representative CCs; this can be quite difficult, but one feasible solution for such an evaluation is to manage all available data as one discrete signal, which allows signal processing tools (SPTs) to be implemented during data analysis. In this research, we present a proof-of-concept study that shows the possibility of mapping to a discrete signal any type of data available from genetic resource collections in order to take advantage of SPTs for the construction of CCs that adequately represent the diversity of two crops. This method is referred to as 'SPT selection.' All available information for each element of the tested collections was analysed under this perspective and compared when possible, with one of the most used algorithms for CC selection. Genotype-only SPT selection did not prove as effective as standard CC selection did not prove as effective as standard CC selection algorithms; however, the SPT approach can consider genotype alongside other types of information, which results in well-represented Ccs that consider both the genotype and agromorphological diversities present in original collections. Furthermore, SPT-based analysis can evaluate all available data both in a comprehensive manner and under different perspective, and despite its limitations, the analysis renders satisfactory results. Thus, SPT-based algorithms for CC selection can be valuable in the field of genetic resources research, management and exploitation.

KEYWORDS: Core Collection, SPT, Genotype, Genebank

RESUMEN

La selección de una colección núcleo (core-collection) representativa (CC) es una estrategia comprobada y eficaz para superar los gastos y las dificultades de la gestión de los recursos genéticos en los bancos de germoplasma de todo el mundo. Debido a las diversas aplicaciones disponibles para estas subcolecciones, se han implementado con éxito varios algoritmos para construirlos en base a datos genotípicos, fenotípicos, de pasaporte o geográficos (ya sea por conjuntos de datos individuales o por consenso). Sin embargo, hasta donde tenemos conocimiento, no se han explorado adecuadamente conjuntos de datos integrales hasta la fecha. Por lo tanto, los investigadores evalúan conjuntos de datos múltiples para construir CCs representativos; esto puede ser bastante difícil, pero una solución factible para tal evaluación es administrar todos los datos disponibles como una señal discreta, que permite implementar herramientas de procesamiento de señal (SPT) durante el análisis de datos. En esta investigación, presentamos un estudio de prueba de concepto que muestra la posibilidad de asignar a una señal discreta cualquier tipo de datos disponibles de colecciones de recursos genéticos para aprovechar los SPT para la construcción de CC que representen adecuadamente la diversidad de dos cultivos. Este método se conoce como "selección de SPT." Toda la información disponible para cada elemento de las colecciones analizadas se analizó bajo esta perspectiva y se comparó cuando fue posible, con uno de los algoritmos más utilizados para la selección de CC. La selección de SPT de solo genotipo no resultó tan efectiva como los algoritmos de selección de CC estándar; sin embargo, el enfoque SPT puede considerar el genotipo junto con otros tipos de información, lo que da como resultado CCs bien representados que consideran tanto el genotipo como las diversidades agromorfológicas presentes en las colecciones originales. Además, el análisis basado en SPT puede evaluar todos los datos disponibles, tanto de manera integral y bajo diferentes perspectivas, y a pesar de sus limitaciones, el análisis arroja resultados satisfactorios. Por lo tanto, los algoritmos basados en SPT para la selección de CC pueden ser valiosos en el campo de la investigación, gestión y explotación de recursos genéticos.

PALABRAS CLAVE: Core Collection, SPT , Banco de germoplasma, genotipo

Correspondencia

DESTINATARIO: **Ernesto Borrayo Carbajal**

INSTITUCIÓN: **Universidad de Guadalajara**

DIRECCIÓN: **Av. Juárez #976, Col. Centro, C.P. 44100, Guadalajara, Jalisco, México**

CORREO ELECTRÓNICO:

ernesto.borrayo@academicos.udg.mx

Fecha de recepción:

27 de septiembre de 2018

Fecha de aceptación:

9 de enero de 2019

INTRODUCTION

One of the most promising techniques for conserving the diversity of genetic resources is *ex situ* genebank germoplasm collection. A significant effort has been made on a global scale to perserve, characterize, distribute and utilise genetic resource in order to understand their biological phenomena and confront the vulnerable situation regarding the sustainability of future human development [1, 2]. As the size of germoplasm colections increase, it becomes difficult to appropriately manage and extensively evaluate them [3]; thus, the core collection (CC) concept [4] has become a fundamental genetic resource management approach and exploits the potential of a complete collection in terms of viable data management and moetary expenses [5, 6, 7, 8].

Different CCs have different purposes characteristics and evaluation criteria [7, 9, 10, 11]; thus, several different algorithms and informatics tools have been developed and implementend [12, 13, 14, 15] with different approaches for satisfying particular needs of each CC. Because these CCs are constructed mainly on the basis of genotypic, phenotypic, passport or geographic data (either by individual datasets or by consensus) [16], there is a lack of all-inclusive datasets; this limits the possibility of generating a CC that may satisfy most basic and applied genetic resource research programs. To the best of our knowledge, no single comprehensive datasets has been properly explored to date.

One possible method to create a comprehensive dataset is to represent the availabe data as numerical values. Several methods exist that represent genomic information into numerical values [17] and agromorphological traits (ATs) into scores [18]. Through this mapping process, treating each data vector as a discrete signal that can, in turn, be analysed by signal processing tools (SPTs) is possible, thus providing and effective tool for a comprehensive evaluation of datasets. We present a proof-of-concept study that shows the possibility of mapping to a discrete signal any type of data available from genetic resource collections in order to take advantage of SPTs for CC selections; this possibility provides new decision-making criteria for genetic resource management and research.

METHODOLOGY

Mapping data

Each input data must be mapped to a numeric value. This is a fundamental process of the algorithm beacause it enable different datasets to be analysed together, regardless of their nature. In this manner, dissimilar passport data, single nuceotide polymorphisms (SNPs), restriction fragment length polymorphisms (RFLPs), geographic information and phenotypical traits can be included in one comprehensive dataset. To consistently represent each data type, reference tables are implemented according to the nature of each particular data: genetic information (originally represented as character elements) is now represented by a numeral vector, and trait variation, simple sequence repeat (SSR) molecular markers and passport data can be represented as either binary or normalized data depending on the quantitative/qualitative nature of the data. The original data and reference tables for this study are available in supplementary material ??. Data transformation for this study rendered a matrix containing the representation of MC samples (i_1, i_2, i_3, ..., i_n) with (j_1, j_2, j_3, ..., j_n) elements each, where n is the total number of samples, and m is the number of included samples characteristics, represented by a numerical values as $data_{(i,j)}$.

Signal construction

Numerical representations of each jth *data* element can be treated as frequency values in m data time in such a manner that each ith sample is treated as a discrete signal. The i signal correspond to the information behaviour from each sample. This perspective will enable the implementation of SPTs such as the dis-

crete Fourier transform and power spectrum comparision. Although SPTs can be implemented on all data available for each sample, not all data elements contain the same informativeness value to discriminate between samples. To overcome the informative difference in each *j* element of *data*, a principal component analysis (PCA) can be perfomed to rearrange *data* into a new matrix that has the high informative elements of *data* at the beginning and that arranges subsequent elements acording to their informativeness, discarding those whose variance equals 0. This process renders two new matrices: the original *characteristics* mapped vectors matrix (*x*) and rearranged variance value matrix (*X*). Matrix *X*, therefore, contains *n* samples that are formed by a numerical vector with *m=m-* (non informative *characteristics*).

Fast Fourier transform

The main objective of Fourier transform is the decomposition of any signal into a complex histogram of frequencies. Signal function is then represented as a vectorial function whose angle and magnitude determine a sampled point in the signal [19].

The original Fourier model is expressed as follows:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi} dx \qquad (1)$$

where *x* is the temporal variable, ξ it the frequential variable, *i* is a -1 square root and *e* is the natural exponent.

From equation (1), a derivate can be determined for any point ξ sampled in the signal.

$$f(x)[\cos^{2\pi e \xi} + i * \sin^{2\pi e \xi}] \qquad (2)$$

Fourier transform can be implemented into any complex numerical series, but in a practical sense, the computational cost increases exponentially.

Thus, fast Fourier transform (FFT) is more often implemented and can be defined according to Cooley-Tukey algorithm [20] as follows:

$$X_k = \sum_{n=0}^{N-1} x_n e^{-i\pi k \frac{n}{N}} \qquad (3)$$

where *N* is the vector length, *x* is the temporal variable, *i* is a -1 square root and *e* is the natural exponent; in such matter that an euclidean representation - with the angle, magnitude and phase that corresponds to their position in the signal - exists for any signal dot.

Therefore, mapping any signal into a vectorial representation that contains information from every original signal dot is possible. From this complex vector, useful data can be retrieved to establish a comparison between them that indirectly represents the original signal's juxtaposition [21].

Distance matrix computation

Inspired by the genomic signal processing alignment-free distance (GAFD) model [22], each signal corresponding to the PCA-mapped accesions data in a set $\hat{S}_i$ was converted into its frequency representation by applying discrete Fourier transform. Its power spectrum F_i was then computed. Subsquently, the distance d(*i,j*) for a given pair of comprehensive data signal was calculated by obtaining the mean square error (MSE) of their respective power spectra:

$$D(i,j) = \sum_x (\hat{F}_i(x) - \hat{F}_j(x))^2 \qquad (4)$$

Finally, a distance matrix (DM) was created by performing a pairwise comparison of all sequences in the set.

In parallel, we construted a point-to-point (RAW) DM on the basis of the MSE given to a pair of signal prior to the PCA analysis.

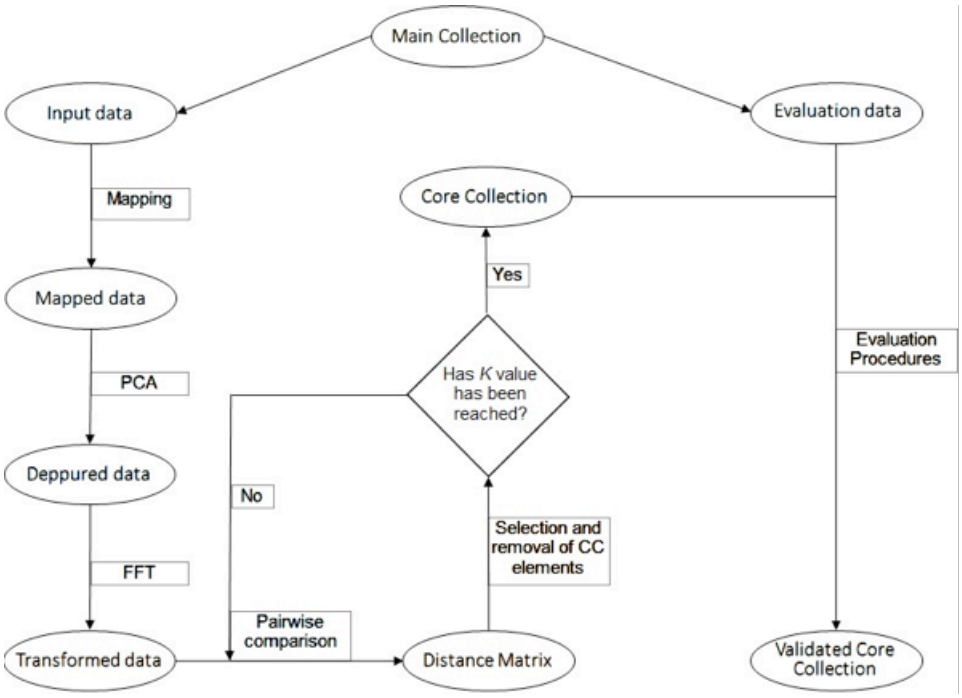


FIGURE 1. General workflow of the FFT-based core collection selection algorithm. PCA: Principal Component Analysis; FFT: Fast Fourier Transfrom; CC: Core Collection.

Core collection selection

Selecting a CC by this method requires the generation of a DM for each sample of the MC; this provides the interrelations among samples and enables adequate selection. A schematic of the complete workflow is present in Fig. 1

In the past, several methodological procedures have been implemented to select K elements from an MC on the basis of information provided by its DM; among such procedures, the most frequently used one is the hierarchical clustering method [11]. However, the current algorithm does not rely on hierarchical clustering for CC selection, instead - similar to the least distance stepwise sampling method [23] - CC elements are selected by an iterative process, where r samples are selected by different criteria (which may be individually implemented) on each iteration.

Selection criteria (based on the MD without hierarchical clustering) for the current algorithm is as follows:

a. The *i*th sample with the most lower distance values among *j*th elements.
b. The *i*th sample with the most higher distance values among *j*th elements.
c. The *i*th sample with a lower distance average.
d. The *i*th sample with a higher distance average.
e. The *i*th sample with a lower overall distance.
f. The *i*th sample with a higher overall distance.

In cases where multiple samples share selection values, an appearance priority will complete the criteria.

An example of selection process is present in Fig. 2 and its final result is present in Fig. 3.

Once the selected samples (r) are included in the future CC, they (along with others that are identical to them (s)) are removed from X for the next iteration; then, a DM_2 with $n_2 = n - r - s$ is calculated. This process will continue Z times until $R >= K$, where $R = (r_1 + r_2 + \ldots + r_Z)$ and K = *predefined* CC *elements desired*.

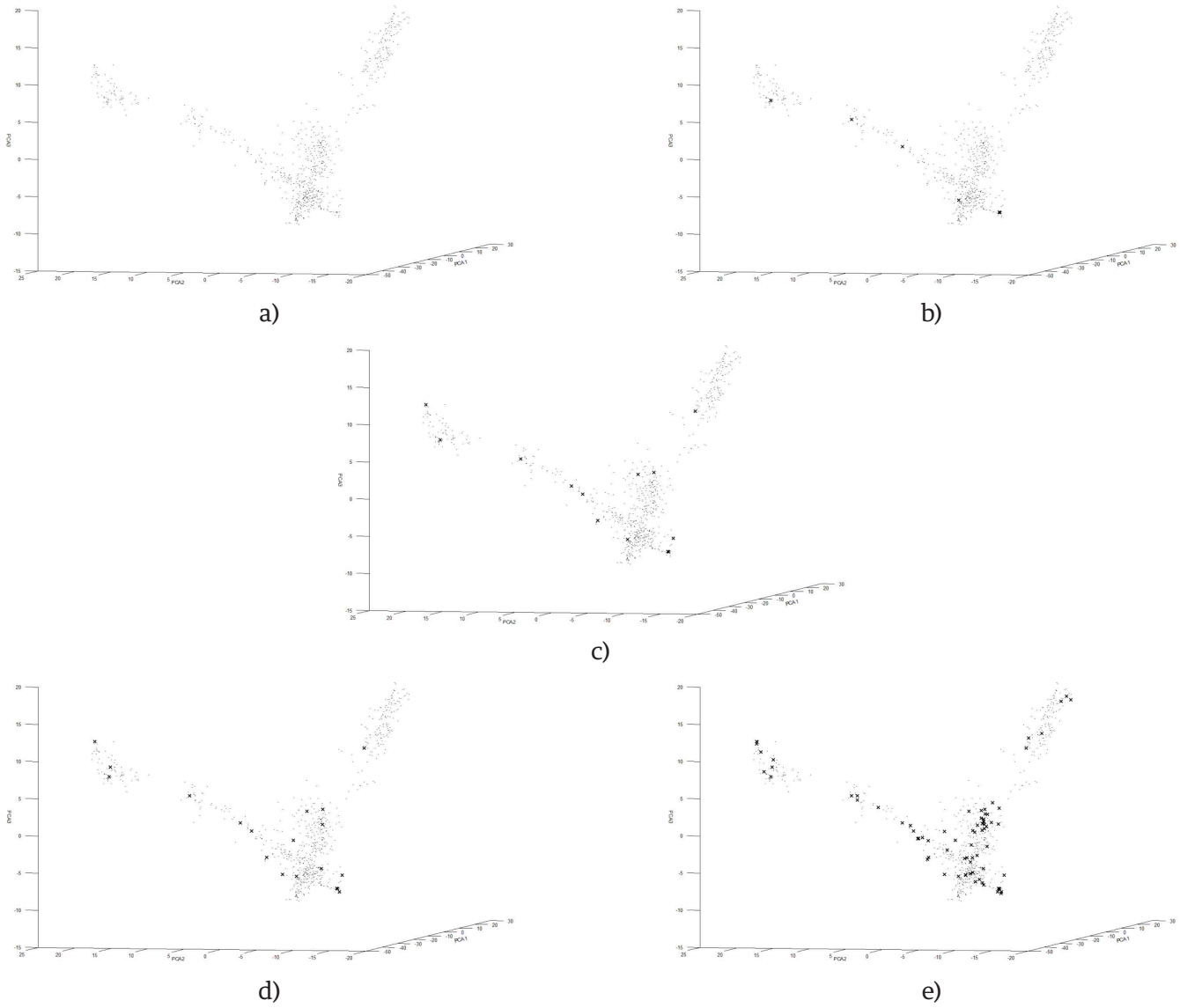


FIGURE 2. **First three principal component's distribution of Rdata (a), methodology's first (b), second (c) and third (d) iteration; final K=72**

Evaluation of the selected core collection

As discussed previously, the best way to evaluate a CC depends on the purpose of that CC, and eve if it can be evaluated from the same dataset from which it was constructed, evaluating it with a diferent dataset [7] is desirable. In this study, we use other datasets for our evaluation whenever possible.The list given below provides the evaluation parameters implemented in this study.

a. The average distance between each MC sample and the nearest CC sample (ANE) can be calculated using the equation as follows:

$$ANE_{tot} = \frac{1}{L}\sum_{k=1}^{K}\sum_{j=1}^{J} D\left(k - cMC_j\right) \qquad (5)$$

where K is all CC elements, k is each CC element and D is the distance between k and each jth cMC

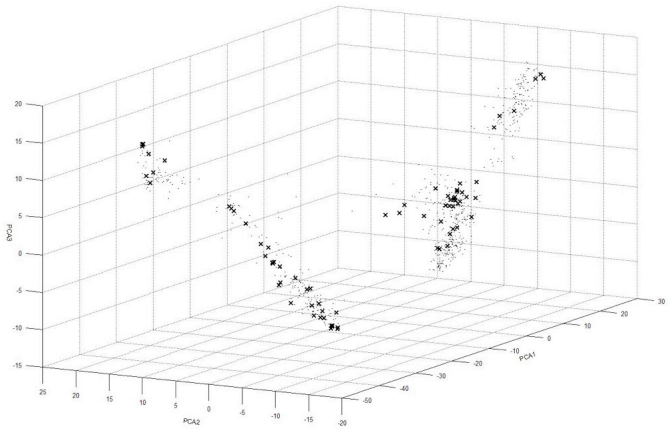


FIGURE 3. **First three principal component's distributions of K=27 CC selection (X) from Rdata MC.**

element whose closest CC element is *k*, including itself, thus rendering *L* total comparisons. The ideal ANE value is 0, where each sample of the CC represents itself and those similar to it. This parameter evaluates the homogeneity of the represented MC diversity.

b. The average distance between each CC sample and the nearest CC sample (ENE) can be calculated using the equation as follows:

$$ENE_{tot} = \frac{1}{L}\sum_{k=1}^{K} D(k - cCC) \qquad (6)$$

where *K* is all CC elements, *k* is each CC element and *D* is the distance between *k* and its closest CC element *cCC*, excluding itself, in *L* total comparisons. With such an evaluation parameter, higher dispersion renders higher scores with the aim of evaluating the dispersion among selected CC elements.

c. The average distance between CC samples (E) can be determined applying the equation as follows:

$$E_{tot} = \frac{1}{L}\sum_{k=1}^{K}\sum_{j=1}^{J} D(k - cCC_j) \qquad (7)$$

where *K* is all CC elements, *k* is each CC element and *D* is the distance between *k* and all other *j*th CC elements *cCC*, excluding itself, in *L* total comparisons. This evaluation parameter indicates higher scores when CC elements have greater distances between themselves.

While previous evaluation parameters are useful for data dispersion analysis, such parameters will not evaluate how well the distribution of the MC is represented on the CC; therefore, the distribution comparisons tests that were included are as follows:

d. The homogeneity test (*F - test* for variances and *t - test* for means; α= 0.05) between the CC and MC for each trait can be represented as a percentage of traits that are statistically different (MD for means and VT for variances) [9].

e. The coincidence rate (CR) can be calculated using the equation as follows:

$$CR = \frac{1}{M}\sum_{m=1}^{M} \frac{R_{CC}}{R_{MC}} \qquad (8)$$

where *R* is the range of each m trait, and *M* represents the number of traits.

f. The variable rate (CV) can be calcuated using the equation as follows:

$$CV = \frac{1}{M}\sum_{m=1}^{M}\frac{CV_{CC}}{CV_{MC}} \quad (9)$$

where *CV* is the coefficient of the variation of each *m* trait in the CC and MC, and *M* is the number of traits. According to Hu *et al*. [10] a valid CC has *CR* > 80 and *MD* <20, which are the limits for the ideal representation of the identity and distribution of the MC.

g. The alleles coverage (CA) can be calculated using the equation as follows:

$$CA = [|1-(|1-ACC|I\,AMC)|] \quad (10)$$

where *ACC* is a set of alleles in the CC, and *AMC* is a set of alleles in the MC; *ACC* meansures the percentage of alleles from the MC that are present in the CC [12].

To compare the obtained CCs with an established methodology, we implemented Core Hunter 2 (CH) [13] as a reference and used it with the program's default parameters on the agrological and genomic datasets.

Experimental datasets

To determine the efficiency of the analysis of data behaviour by point-to-point direct comparison, a synthetic dataset esa constructed using binary data (*Sdata*) with manageable *n* and *m* elements .

To test the algorithm in real biological-context scenarios, the CCs from different Mcs were constructed and evaluated.

To test the algorithm's CCs versus the scores of the MCs, 780 rice (*Oriza sativa* (L.)) accession and 423 foxtail millet (*Setaria italica subspitalica* (L.) *P. Beauv.*) accession data were retrieved from the then National Institute of Agrobiological Sciences (now National Agriculture and Food Research Organization [NARO]) *http://www.gene.affrc.go.jp/databases_en.php* as well as 361 maize (*Zea mays* (L.)) from the International Maize and Wheat Improvement Center public repository.

According to the available data, different datasets were assembled. The 762 SNPs from the 780 rice accesion retrieved from the NARO database (*Rdata*) were divided arbitrarly into two subsets of 331 SNPs each for constructing two smaller datasets (RdataI and *RdataIII*). In addition, ATs were categorized and mapped into the binary data for 273 of the 780 accessions, resulting in 38 variables (*RdataII*). The variables from 423 foxtail millet genotypes with transposon displays [24] were used as a single dataset (*Fdata*). For a subset of 141 accessions (*FdataI*), 9 ATs were categorized and mapped into binary data, resulting in 28 variables (*FdataII*). The maize available information was mapped into 0-1 values (*Mdata*). The substitution tables used during this mapping are presented as supplementary material 1.

Implementation

All procedures were implemented in python 3.6, codes are available as supplementary material 2.

A graphical interface was developed including a SQLite3 database (*https://sqlitebrowser.org/*) in order to store data for future comparison and further analysis. This implementation includes a previously described K-means based CC selection algorithm [25].

RESULTS AND DISCUSSION

Selection and evaluation

The selection criteria were chosen to look for the best possible distribution of selected CC elements within the DM. Although hierarchical clustering has proven to be an effective method for determining collection

structure and sampling CC [26] and although it has been implemented in different crop [27, 28] and included in various selection algortihms [11], hierarchical reconstruction presents the challenge of selecting an appropriate model for biological interpretation that can be applied to everything from unweighted pair-group averages to Markov models in Bayesian estimations [29]. To avoid the challenge of selecting a reconstruction model, we decided to work strictly with the DM. By selecting the items described in this methodology, we aimed to retrieve representative elements from among the distributions of collections; however, because of its iterative nature, this methodology may render high redundancy under certain data distributions. Despite this limitation, the methodology has proven to be capable of selecting representative elements of the MC's diversity.

Evaluation criteria were applied according to Odong *et al*. [7] without excluding the classic criteria used in [9, 10]. The selected CCs render proper results in general terms. As expected, selected CCs did not always reach for optimal values for MD and CR, this is due the fact that it is not the aim of the selection method to render a CC with similar distribution to that of the MC, but to make sure to include as much diversity as possible.

It is our belief that scoring the CC sets obtained with these methodologies will enable genetic resource banks to provide clear descriptors of what their CC strengths and limitations are with respect to the MC from which they come and will provide adequate tools for determining the possible purposes of the selected CCs.

Although several representations of genotypic characteristics (particulary those involving DNA sequences [30, 31, 32]) have been proposed, real-number-based mappings have not been discarded, indeed, this type of mapping has been highly studied for signal analysis even when they share two principal problems: the preferential magnitude of some nucleotides and the non-equidistance of all nucleotides [33, 34]. The arbitrary values selected for SNP's numerical representation of genotypes aim to maintain equidistance relations among purines and among pyrimidines in such a manner that the same distance is also perserved between at least one of them and the undetermined values. ATs are represented as binary data. This representation may prove useful for discrete data but requires a clustering procedure for continuous data. In this study, we arbitrarily generated clusters for the latter and then represented them as the former. Although this implementation may no be the most accurate regarding biological or agronomical significance, it serves as the first approach for testing the feasibility of the use of signal processing techniques when merging several datasets to construct one CC.

RAW versus FFT

The RAW comparison establishes a distance value on the basis of the average distance between each mapped value on each element while the FFT power spectra implementation compares the signals in the frequency domain. Using FFT, establishing a DM on the basis of how data 'shifted' rather than on the basis of average point-to-point comparisons was possible. The FFT approach provides a different DM, where its compared elements are clustered based on the similarity of the shift is in the opposite phase. We expect that the procedure reveal more info about the relations between the individual components within each element.

FFT comparisons of signal without PCA are a good approach for CC selection. Nervertheless, PCA implementation enables us to avoid possible misleads in random data arrangements, as, for example, palindromic data that could result in the same power spectra. Moreover, through PCA, we could organize data according to their levels of impact on the difference between accessions, which --when their magnitudes were obtained-- inherently rendered a representation

TABLE 1. K selected CC scores from MC Sdata Raw and PCA Signal evaluated with Sdata

	Sdata PCA			Sdata RAW		
K	12	18	24	12	18	24
ANE	0.2348	0.2311	0.2164	0.2697	0.2287	0.2164
ENE	0.339	0.3386	0.3401	0.3696	0.3228	0.3214
E	0.5562	0.5622	0.5547	0.5558	0.5333	0.5299
MD	0	0	0	0	0	0
VT	41.6667	50	41.6667	33.3333	58.3333	41.6667
CR	64.8403	71.6918	73.7154	60.6447	75.2465	80.4716
CV	9080.798	61.2074	86.0876	136.6446	139.1418	280.8481
AR	74.3363	81.4159	89.3805	61.9469	77.8761	80.531

TABLE 2. K selected CC scores from MC Fdata Raw and PCA signal evaluate with Fdata

	Edata PCA			Edata RAW		
K	48	72	96	48	72	96
ANE	0.6454	0.6423	0.6407	0.6489	0.6431	0.643
ENE	0.646	0.6472	0.6472	0.65	0.6448	0.6452
E	0.7297	0.7301	0.7301	0.7231	0.7236	0.7239
MD	1.1799	0.59	0.59	1.7699	1.4749	1.4749
VT	50.4425	53.6873	56.6372	50.7375	56.0472	55.1622
CR	83.6883	87.0605	88.9709	83.5334	86.9308	87.7461
CV	0.8494	0.419	0.7357	1.1037	4.74	0.7361
VA	96.3945	97.7652	98.5995	95.3516	97.497	97.4374

TABLE 3. K selected CC scores from MC Rdata Raw and PCA Signal evaluated with Rdata

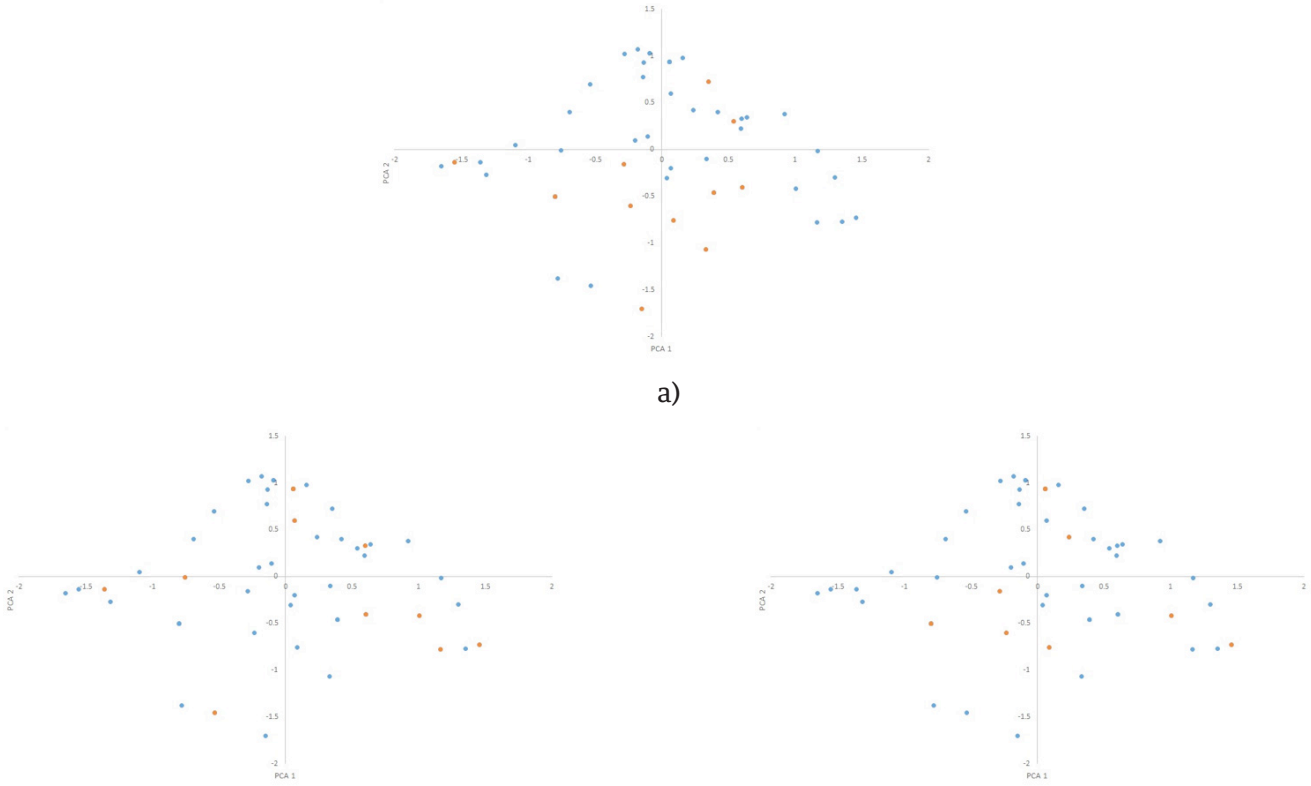
	Rdata PCA			Rdata RAW		
K	48	96	156	48	96	156
ANE	0.6013	0.5966	0.5942	0.6118	0.6052	0.6042
ENE	0.5939	0.5944	0.5981	0.6106	0.6085	0.609
E	0.7105	0.7074	0.7051	0.703	0.7038	0.7054
MD	9.1146	5.9896	3.9062	10.1562	5.4688	4.4271
VT	42.4479	48.6979	58.0729	57.5521	72.9167	70.0521
CR	70.5716	78.477	83.2957	69.9022	78.1045	80.0167
CV	1.0171	0.4343	0.3137	7.9407	0.4375	1.1344
VA	92.6758	96.8992	98.5298	93.9856	98.1823	98.5031

of informativity relations among values. This 'data behaviour' was used as the element for pairwise comparisons, and although this approach clusters differently from RAW comparisons, we belive that it will provide a new perspective for CC selection and open the possibility of further data exploration.

Our first approach was to measure the comparisons under different *K* values. We compared the approach of the RAW signals whit the PCA-FFT- treated signals. Results from *Sdata*, *Fdata*, and *Rdata* are presented in Tables 1-3. As expected most evaluation criteria improved as *K* increased.

The use of FFT signals renders better overall scores than use of RAW signal in *Sdata* and *Fdata*; however, this advantage diminishes in *Rdata*. We speculate that this difference can be explained by the mapping procedures used; further research regarding this matter is ecouraged.

Using the CH's rendered *K* values, we used both CH and FFT to generate the CCs is summarized in Table 4 and in Figs 4-5. Both methodologies rendered similar results, yet PCA rendered better results on parameters representing MC distribution; this could be an effect of the selection method's intrinsic redundancy.



a) b) c)

FIGURE 4. First two principal component's distributions of k=11 CC (orange) selected by CH(a), PCA(b) and RAW (c) in Sdata distribution (blue).

a) b)

FIGURE 5. First two principal component's dristibutions of k=84 CC (orange) selected by CH(a) and PCA(b) in Fdata distribution (blue).

To further this concept, we analized maize data with both K-means and FFT implementation, in order to both contrast with a differt approach and test the interfase. The results are presented in Fig 6.

Thus far, the proposed CC selection method and algorithm appear worthy of further exploration. We are aware that two particular fundamental elements require immediate attention. First, a better mapping

TABLE 4. CCs selected from MC Sdata, Fdata and Rdata using PCA signals and Core Hunter compared with respective same data

	Sdata		Fdata		Rdata	
	PCA CH		PCA CH		PCA CH	
K	12		84		156	
ANE	0.2348	0.2314	0.6407	0.6392	0.5942	0.5952
ENE	0.339	0.3906	0.6474	0.6386	0.5981	0.6047
E	0.5562	0.563	0.7304	0.7176	0.7051	0.7017
MD	0	0	0.59	1.1799	3.9062	5.4688
VT	41.6667	58.3333	56.6372	66.6667	58.0729	86.7188
CR	65.6045	76.1001	88.9709	93.0119	83.2957	89.6723
CV	9080.978	132.6078	0.7357	0.429	0.3137	0.4001
AR	74.3363	76.9912	98.5995	98.4803	98.5298	99.3852

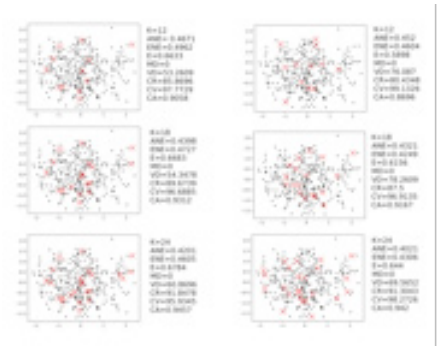


FIGURE 6. First two principal component's distributions of k=12 CC (up), k=18 CC (center) and k=24 CC (bottom); selected by FFT (left) and K-means (right) with their respective evaluation values. Black dots correspond to the complete maize set, while red X represent selected elements for CC.

solution for both genotypic and AT numerical representation needs to be determined. Second, the selection system developed by us is directly based on the DM and is prone to high redundancy in some data distributions. As discussed earlier, this selection system was chosen in order to avoid the problems associated with hierarchical clustering and further allocation selections [13, 35]. Both issues should be addressed in the near future.

Comprehensive data analysis

To demonstrate thet FFT-based CC selection can include and analyse data regardless of its origin, we concatenated corresponding signals from FdataI whit FdataII as well as RdataI and RdaraIII with RdataII to construct Mfdata, MRdataI and MRdataIII. The comprehensive sts were used to construct CCs; the sets were then compared with both their orginal genotype and phenotype MCs. These comparisons are shown in Tables 5-8, and their distributions are represented in Fig. 7-10.

These comprehensive CCs showed overall better scores than genotypic-only CCs when compared with genotypic-only data. On the contrary, there was a better overall score in phenotypic-only CCs when compared against phenotypic-only data.

In the latter case, it should be kept in mind that comprehensive data also consider genotypic data; this could explain why better selections are made when only phenotypic data are considered because genotypic variations may reduce the impact of some phenotypic traits in the PCA analysis.

The generation of a DM based on signal comparisons originating from mixed data construction enables us to explore one of the most interesting applications of this algorithm. By mapping genotypic and AT data, constructing a single signal with all data avaliable for a particular accession is possible. The possibility of including genotypic data with phenotypic traits, geographical locations, climates, habitats, nutritional requirements, symbiotic relationships and so forth provides an opportunity for determining the best information to be included in the selection process in order to cope with the particular objectives for which that CC is beging selected. This concept, in addition to adequate scoring systems, may prove useful in designing tailored CCs that comply with specific research/breeding objetive.

TABLE 5. CCs selected from MC FdataI and MC MFdata PCA signals and evaluated with FdataI and FdataII

	vs EdataI		vs EdataII	
	FdataI	MFdata	FdataII	MFdata
K	24			
ANE	0.6333	0.6356	0.4049	0.4093
ENE	0.6413	0.6423	0.4374	0.4351
E	0.7194	0.7113	0.623	0.5914
MD	1.7668	2.4735	0	0
VT	66.0777	33.9223	46.42	64.2857
CR	89.4908	89.8198	80.677	82.1913
CV	45.7033	35.6847	21.8658	132.1517
AR	91.7647	92.7206	97.5904	94.3775

TABLE 6. CCs selected from MC RdataI , MrdataI, RdataIII and MRdataIII PCA signals and evaluated with RdataI

vs RData				
	RdataI	MRdataI	RdataIII	MRdataIII
K	24			
ANE	0.6148	0.6156	0.6251	0.6169
ENE	0.5989	0.6107	0.621	0.6194
E	0.6962	0.6909	0.6985	0.6934
MD	8.8542	8.5938	7.2917	6.7708
VT	52.0833	63.5417	52.0833	53.3854
CR	80.7367	83.768	81.7278	81.8623
CV	56.3949	59.6279	45.6875	199.9377
AR	86.5097	88.144	86.5651	90.7202

TABLE 7. CCs selected from MC RdataI, MRdataI, RdataIII and MRdataIII PCA signals and evaluated with RdataIII

vs RDataIII				
	RdataI	MRdataI	RdataIII	MRdataIII
K	24			
ANE	0.6285	0.6276	0.6314	0.623
ENE	0.6273	0.6294	0.6368	0.6267
E	0.7036	0.7054	0.7226	0.7056
MD	8.0729	7.5521	7.2917	10.4167
VT	52.8646	60.6771	51.5625	46.875
CR	79.5995	81.0356	79.6809	84.53
CV	28.3673	56.3689	90.0475	60.7279
AR	88.9071	88.7705	87.5956	93.0471

TABLE 8. CCs selected from MC FdataI and MC MFdata PCA signals and evaluated with FdataI and FdataII

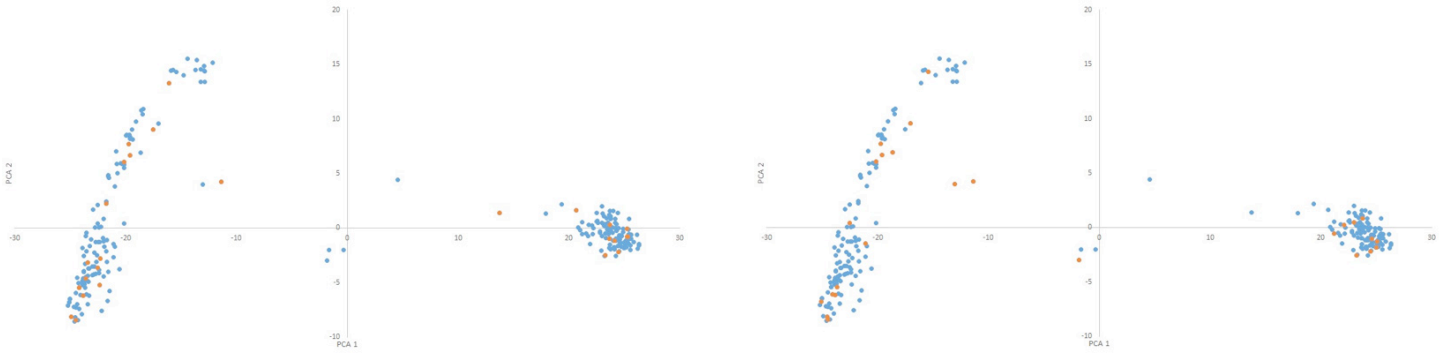
vs RDataII			
	RdataII	MRdataI	MRdataIII
K	24		
ANE	0.4594	0.4652	0.4618
ENE	0.4796	0.4896	0.4742
E	0.6402	0.6205	0.6169
MD	0	5.2632	0
VT	39.4737	42.1053	60.5263
CR	63.8082	61.8988	68.2437
CV	3.8262	2.2285	4.1332
AR	95.4268	98.7805	98.7805

a) b)

FIGURE 7. First two principal component's distributions of k=24 CC (orange) selected by PCA from Fdata(a) and Mdata(b) in FdataI distribution (blue).

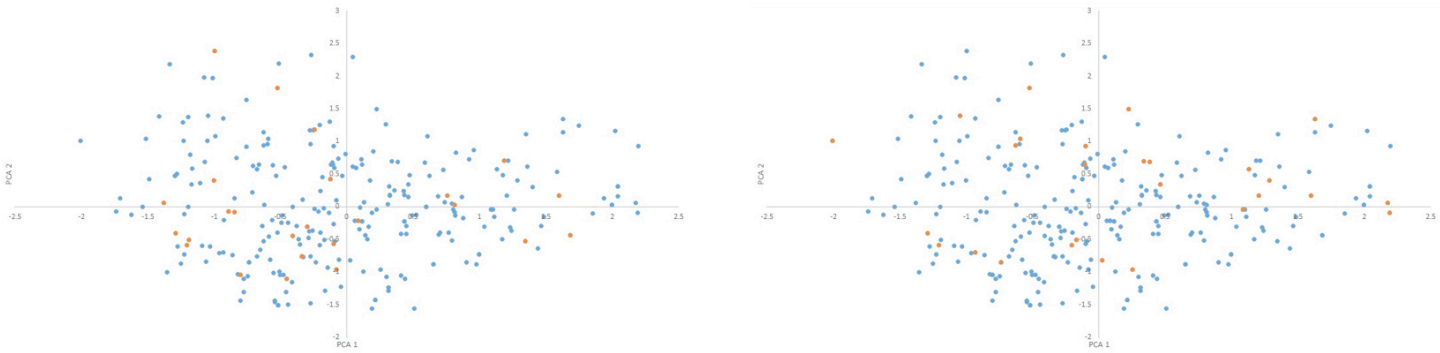
a) b)

FIGURE 8. First two principal component's distributions of k=24 CC (orange) selected by PCA from FdataII(a) and Mdata(b) in FdataII.



a) b)

FIGURE 9. First two principal component's distributions of k=24 CC (orange) selected by PCA from RdataIII (a) and MdataIII (b) in RdataI distribution (blue).



a) b)

FIGURE 10. First two principal component's distributions of k=24 CC (orange) selected by PCA from RdataII (a) and MRDataI (b) in RdataII distribution (blue).

CONCLUSIONS

The use of SPTs in CC selection, as presented in this algorithm, enables us to analyse all avaliable data comprehensively and from different perspectives. Despite its limitations, this signal construction make it possible to analyse all avaliable data regarding each accesion in CC selection with good results.

The efficiency of SPTs in CC selection sugests that the use of these tools in MC analysis may provide useful information not only for CC but also for other purposes.

The implementation of current and other SPTs in all-inclusive MC-mapped signals is worth further exploration, and we belive that it will be an important asset to genentic resource management and exploitation.

AUTHOR CONTRIBUTIONS

ILF performed the implementation, helped with the analysis and manuscript drafting. MT contributed to the design of the algorithm, data analysis and manuscript drafting and correction. EB conceived and designed the algorithm, performed the implementation, analysed the data and wrote the manuscript. All authors have read and approved the final manuscript.

COMPETING INTERESTS

No competing interests were disclosed.

GRANT INFORMATION

This research is supported in part by the SATREPS project by JST and JICA, Diversity Assessment and Development of Sustainable Use of Mexican Genetic Resources and in part by JSPS Grant-in-Aid 25257416.

REFERENCIAS

[1] Biotechnology, P., & Watanabe, K. N. (1999). Plant Genetic Resources and its Global, 7-13.

[2] Dulloo M, Hunter D, Borelli T. Ex situ and in situ conservation of agricultural biodiversity: major advances and research needs. Not Bot Horti ... [Internet]. 2010 [cited 2014 Nov 19];38(2):123-35. Available from: *http://notulaebotanicae.ro/index.php/nbha/article/view/4878*

[3] Upadhyaya H, Gowda C, Sastry D. Plant genetic resources management: collection, characterization, conservation and utilization. J SAT Agric ... [Internet]. 2008 [cited 2014 Nov 19];6(December):1-16. https://doi.org/10.1186/1471-2229-8-106

[4] Brown, A. H. D. (1989). Core collections: a practical approach to genetic resources management. Genome, 31(2), 818-824. https://doi.org/10.1139/g89-144

[5] Guo Y, Li Y, Hong H, Qiu L-J. Establishment of the integrated applied core collection and its comparison with mini core collection in soybean (Glycine max). Crop J [Internet]. 2014 Feb [cited 2014 Jun 6];2(1):38-45. Available from: *http://linkinghub.elsevier.com/retrieve/pii/S2214514113000366*

[6] Studnicki M, MADRY W, Schmidt J. Efficiency of Sampling Strategies to Establish a Representative in the Phenotypic-based Genetic Diversity Core Collection of Orchardgrass (Dactylis glomerata. Czech J Genet Plant Breed [Internet]. 2013 [cited 2014 Jul 10];2013(1):36-47. Retrieved from: *https://goo.gl/vfGhYu*

[7] Odong, T. L., Jansen, J., van Eeuwijk, F. A., & van Hintum, T. J. L. (2013). Quality of core collections for effective utilisation of genetic resources review, discussion and interpretation. Theoretical and Applied Genetics, 126(2), 289-305. https://doi.org/10.1007/s00122-012-1971-y

[8] Richards C, Volk G. Selection of stratified core sets representing wild apple (Malus sieversii). J Am Soc Hortic Sci [Internet]. 2009 [cited 2014 Jul 31];134(2):228-35. Retrieved from: *http://journal.ashspublications.org/content/134/2/228.short*

[9] Franco J, Crossa J, Warburton ML, Taba S. Sampling Strategies for Conserving Maize Diversity When Forming Core Subsets Using Genetic Markers. Crop Sci [Internet]. 2006 [cited 2014 Jun 19];46(2):854. https://doi.org/10.2135/cropsci2005.07-0201

[10] Hu J, Zhu J, Xu HM. Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. TAG Theor Appl Genet [Internet]. 2000 Jul 12;101(1-2):264-8. https://doi.org/10.1007/s001220051478

[11] Wang J, Hu J, Huang X, Xu S. Assessment of different genetic distances in constructing cotton core subset by genotypic values. J Zhejiang Univ Sci B [Internet]. 2008 May [cited 2014 Jul 1];9(5):356-62. Available from: *http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2367373&tool=pmcentrez&rendertype=abstract*

[12] Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, Davenport GF. Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics [Internet]. 2009 Jan [cited 2014 Jun 19];10:243. https://doi.org/10.1186/1471-2105-10-243

[13] De Beukelaer H, Smýkal P, Davenport GF, Fack V. Core Hunter II: fast core subset selection based on multiple genetic diversity measures using Mixed Replica search. BMC Bioinformatics [Internet]. 2012 Jan;13:312. https://doi.org/10.1186/1471-2105-13-31

[14] Jansen J, van Hintum T. Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. Theor Appl Genet [Internet]. 2007 Mar [cited 2014 May 29];114(3):421-8. https://doi.org/10.1007/s00122-012-1971-y

[15] Gouesnard B, Bataillon T. MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. J ... [Internet]. 2001 [cited 2014 Jul 1];93-4. Retrieved from *http://jhered.oxfordjournals.org/content/92/1/93.short*

[16] Franco J, Crossa J, Ribaut J. A method for combining molecular markers and phenotypic attributes for classifying plant genotypes. Theor Appl Genet [Internet]. 2001 [cited 2014 Aug 18];103:944-52. Retrieved from *https://goo.gl/H84Nc8*

[17] Kwan HK, Arniker SB. Numerical representation of DNA sequences. Proc 2009 IEEE Int Conf Electro/Information Technol EIT 2009. 2009;307-10. https://doi.org/10.1109/EIT.2009.5189632

[18] Dossou-aminon I, Loko LY, Adjatin A, Ewédjè EBK, Dansi A, Rakshit S, et al. Genetic Divergence in Northern Benin Sorghum (Sorghum bicolor L . Moench) Landraces as Revealed by Agromorphological Traits and Selection of Candidate Genotypes. Sci World J. 2015;2015:e916476. https://doi.org/10.1155/2015/916476

[19] Stein E, Weiss G. The Fourier Transform. In: Introduction to Fourier analysis on Euclidean Spaces. Princeton University Press; 1971.

[20] Cooley J, Tukey J. An algorithm for the machine calculation of complex Fourier series. Math Comput [Internet]. 1965 [cited 2012 Nov 10];297-301. https://doi.org/10.2307/2003354

[21] Nagarajan N, Keich U. FAST: Fourier transform based algorithms for significance testing of ungapped multiple alignments. Bioinformatics [Internet]. 2008 Feb [cited 2011 Sep 17];24(4):577-8. https://doi.org/10.1093/bioinformatics/btm594

[22] Borrayo, E., Mendizabal-Ruiz, E. G., Velez-Perez, H., Romo-Vazquez, R., Mendizabal, A. P., & Alejandro Morales, J. (2014). Genomic signal processing methods for computation of alignment-free distances from dna sequences. PLoS ONE, 9(11), 1-13. https://doi.org/10.1371/journal.pone.0110954

[23] Wang J, Guan Y, Wang Y, Zhu L, Wang Q, Hu Q, et al. A strategy for finding the optimal scale of plant core collection based on Monte Carlo simulation. ScientificWorldJournal [Internet]. 2014 Jan;2014:503473. Available from: *https://goo.gl/gA8LsF*

[24] Hirano R, Naito K, Fukunaga K. Genetic structure of landraces in foxtail millet (Setaria italica (L.) P. Beauv.) revealed with transposon display and interpretation to crop evolution of foxtail millet. ... [Internet]. 2011 [cited 2014 Jul 10];506:498-506. https://doi.org/10.1139/g11-015

[25] Borrayo E, Machida-Hirano R, Takeya M, Kawase M, Watanabe K. Principal components analysis - K-means transposon element based foxtail millet core collection selection method. BMC Genet [Internet]. 2016 Dec 16;17(1):42. https://doi.org/10.1186/s12863-016-0343-z

[26] Odong TL, van Heerwaarden J, Jansen J, van Hintum TJL, van Eeuwijk F a. Determination of genetic structure of germplasm collections: are traditional hierarchical clustering methods appropriate for molecular marker data? Theor Appl Genet [Internet]. 2011 Jul [cited 2014 Jan 21];123(2):195-205. https://doi.org/10.2135/cropsci2011.02.0095

[27] Cericola F, Portis E, Toppino L, Barchi L, Acciarri N, Ciriaci T, et al. The population structure and diversity of eggplant from Asia and the Mediterranean Basin. PLoS One [Internet]. 2013 Jan [cited 2014 Jun 28];8(9):e73702. https://doi.org/10.1371/journal.pone.0073702 ¿?

[28] Mei Y, Zhou J, Xu H, Zhu S. Development of sea island cotton ('Gossypium barbadense'L.) Core collection using genotypic values. Aust J Crop Sicence [Internet]. 2012 [cited 2014 Jul 3];6(4):673-80. Retrieved from: *http://search.informit.com.au/documentSummary;dn=362661761803357;res=IELHSS*

[29] Redelings BD, Suchard MA. Joint Bayesian estimation of alignment and phylogeny. Syst Biol [Internet]. 2005;54(3):401. https://doi.org/10.1080/10635150590947041

[30] Cristea, P. D. (2002). Conversion of nucleotides sequences into genomic signals. Journal of Cellular and Molecular Medicine, 6(2), 279-303. https://doi.org/10.1111/j.1582-4934.2002.tb00196.x

[31] Rosen GL, Sokhansanj B a, Polikar R, Bruns MA, Russell J, Garbarine E, et al. Signal processing for metagenomics: extracting information from the soup. Curr Genomics [Internet]. 2009 Nov [cited 2012 Mar 29];10(7):493-510. https://doi.org/10.2174/138920209789208255

[32] Wang LWL, Schonfeld D. Mapping Equivalence for Symbolic Sequences: Theory and Applications. IEEE Trans Signal Process [Internet]. 2009;57(12):4895-905. https://doi.org/10.1109/TSP.2009.2026544

[33] Almeida JS, Vinga S. Universal sequence map (USM) of arbitrary discrete sequences. BMC Bioinformatics [Internet]. 2002;3(1):6. Retrieved from: *https://www.dropbox.com/s/0w3fllpy9ln250g/Almeida_etal_2002_JZSER_LMO.pdf*

[34] Akhtar M, Epps J, Ambikairajah E. Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction. IEEE J Sel Top Signal Process [Internet]. 2008 Jun;2(3):310-21. https://doi.org/10.1109/JSTSP.2008.923854

[35] Franco J, Crossa J, Taba S, Shands H. A Sampling Strategy for Conserving Genetic Diversity when Forming Core Subsets. Crop Sci [Internet]. 2005 [cited 2014 Dec 4];45(3):1035. https://doi.org/10.2135/cropsci2004.0292