Australian Journal of Crop Science AJCS

AJCS 16(05):580-586 (2022) ISSN:1835-2707
doi: 10.21475/ajcs.22.16.05.p3534

Metrics for optimum allocation of resources on the composition and characterization of crop collections: The CIMMYT wheat collection as a proof of concept

M. Humberto Reyes-Valdés[1], Juan Burgueño[2]*, Carolina Paola Sansaloni[2], Thomas Payne[2], Angela Pacheco[2] and Areli González-Cortés[3]

[1]Department of Plant Breeding, Universidad Autónoma Agraria Antonio Narro, Saltillo, Coahuila, México,
[2]International Maize and Wheat Improvement Center (CIMMYT), México City, México
[3]Instituto de Ciencias y Humanidades Lic. Salvador González Lobo, Universidad Autónoma de Coahuila, México

***Corresponding author: mathgenome@gmail.com**

Abstract

Crop genebank collections are important resources for preserving genetic diversity to face the worldwide demand for food and coping with crop diseases and climate change. However, genebanks tend to accumulate materials without systematic collection growth. Thus, tools for optimizing collections are expected to help improvement of genebanks quality. Furthermore, the genotyping efforts of genebanks would benefit from tools that can help to sample the accessions. A set of parameters to aid the optimization of genebanks are defined, in which Relative Balance is central. In this study, the foundation of our mathematical approach was Kullback-Leibler divergence, providing formulas with consistent properties. Two examples were used as proof of concept. The first one was the comparison between actual and putative optimal numbers of accessions in the Triticum set of the CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo) Wheat Germplasm Bank, with 135,236 entries classified into ten groups. The second one was based on a set containing Triticum plus eight related genera, with 159,741 accessions classified into 217 end-groups, with the goal of illustrating the use of the analytical tools to optimize the ongoing genotyping process. The first example shows a scenario with a well-balanced allocation of accessions. The second example illustrates the optimized choice of end-groups to add 10,000 accessions to the genotyping process. The proof of concept showed the consistency and usefulness of the proposed methods for the improvement of composition in collections and their characterization.

Keywords: Crop genebanks; Optimization; Relative Balance; Wheat.
Abbreviations: CIMMYT_Centro Internacional de Mejoramiento de Maíz y Trigo; CWR_ Crop wild relatives; CCII_ Collection composition imbalance index; RB_ Relative balance; BEUROPE_ Balkans and Eastern Europe; CWANA_ Central and West Asia and North Africa; WEUROPE_ Western Europe; OCEANIA _ Oceania countries excluding America; OTHER_ Remaining world countries; D_ Divergence.

Introduction

The aim of crop genebanks is to conserve and access of plant genetic diversity to researchers and breeders. Usually banks tend to accumulate materials, without considering balancing diversity conservation with stakeholder and client needs, and costs of operations. An important goal for the optimization of a collection is to understand the relationship between the collection composition and the diversity required to meet user needs. In the last years, the scientific community has put more emphasis on considering the conservation of wild relatives, in situ collection, and the role of stakeholders and in situ collections (Frese et al., 2014; Radu-Liviu Sumalan et al., 2021; Singh et al., 2019). However, conceptual tools for collection optimization are scarce in the scientific literature (Van Treuren et al., 2009). Failure to follow a systematic growth of genebanks may lead to under-representation of neglected and underutilized genepools, species, and crop wild relatives (Engels et al., 1995). Crop wild relatives are now being regarded as an increasingly important component of seed banks, although they are still poorly represented (Maxted et al., 2012; Smale and Jamora, 2020).

The advent of high-throughput genotyping techniques is steadily increasing genetic data available for genebank accessions. This information may be used to understand diversity, identify gaps, and reduce redundancy between accessions. Molecular genetic information can be used to develop core subsets of materials within genebank collections, aimed to maximize various criteria of genetic diversity. Acuña-Matamoros and Reyes-Valdés (2018) described and compared several core subset methods, while Reyes-Valdés et al. (2018) proposed a method based on information theory.

Molecular markers provide unprecedented amounts of information for many applications related to genebanks, like genome-wide association studies, genomics based prediction, selection footprints, and genetic diversity studies. However, it

has been argued that molecular markers in the short term are more likely to impact such user-oriented activities, rather than the operation of genebanks. Molecular techniques appear to be mostly a component of the services provided by genebanks, apart from the traditional conservation tasks and sharing of seeds. Molecular information of the accessions provides a way to optimize the user-oriented services offered by the genebanks (Van Treuren and Van Hintum, 2014).

Crop genebanks do not escape from resource limitations. They require adequate installations, space, electrical energy, and above all, human efforts for the maintenance of their genebank collections. Moreover, viability monitoring, genetic integrity, phenotyping and genotyping are activities that are essential for a vibrant collection. Genebank management programs should try to acquire, maintain, distribute, preserve, characterize, evaluate and enhance genetic diversity rather than simply acquiring and storing accessions (Goodman, 1990; Shands, 1990). Thus, decisions should be made on what sorts of accessions must be sought, retained, or discarded. Climate change adds an uncertainty factor in which conserved genetic diversity will be necessary for crop improvement. Crop wild relatives (CWR) will increasingly become a critical resource for crop improvement leading to the sustainability of global food security (Maxted et al., 2012; Vincent et al., 2013). De Oliveira Silva et al. (2019) called the attention to different biobank collections with a specialized productive purpose. Therefore, some collections emphasize indigenous and cultural breed attributes while others are held as public good resources in networks.

Thus, systematization is a *sine qua non* for seed banks, and it has been well attended with standards for germplasm handling, characterization methods, evaluation, documentation, exchange and personnel security (FAO, 2014). However, there is a paucity of standards and little agreement on genebank composition requirements. Quantitative indexes for composition definition and strategic dynamics leading to improved balance, coverage and optimization of genotyping efforts will help to fill this void. The same concepts developed to create core collections (Frankel and Brown, 1984; Brown, 1989b) can be used not only to improve genebank management and utilization management (Brown, 1989a), but also to optimize the complete collection as proposed by Laghetti et al. (2008), as well as the genotyping process.

Currently, the composition of genebanks is often based on the local or national interests, sharing of materials between banks, and serendipity. Although the composition of genebanks may have been managed by curators mainly based on individual experiences, acquisition opportunities, available resources and users demand should be criteria that systematically drive the composition of collections (Van Treuren et al., 2009). Van Treuren et al. (2009) further proposed a strategy to logically define a collection largely based on the concept of core collections, which rely on a hierarchical description of the components of a genebank, weighted by relative priorities. Thus far, there are few approaches to objectively determine those proportions, with reliance on subjective assessments by bank curators and specialists. To facilitate the optimization of such a hierarchical structure, they proposed a composition imbalance index that describes the difference between an actual and an ideal genebank. Nonetheless, the proposed metrics require the definition of optimal proportions for the different groups stored in a genebank. The process to optimize the genebank collections considers two steps: i) define the diversity tree as proposed by Van Treuren (2009), and ii) conduct an optimal allocation of samples to subgroups to achieve the optional proportions of the diversity tree. Although such an index is a useful advance for a systematic optimization of a collection of genetic resources, it may give inconsistent results. In our research, we propose an index with consistent mathematical properties, based on information theory. Such an index can be used, given an optimal composition of the genebank, in an optimization strategy for either the composition of the genebank or the genotyping coverage. As a proof of concept, we apply the methods to the CIMMYT Wheat Germplasm Collection. Information theory tools have been previously used for core subset estimation using molecular marker data, with the introduction of novel concepts of accession rarity and divergence, as well as marker allele specificity (Reyes-Valdés et al., 2018).

Results

Relative Balance and Coverage

In the Materials and Methods section we discuss some drawbacks of the composition imbalance index (CCII) proposed by Van Treuren et al. (2009). Then, we develop the concept of Relative Balance, based on the Kullback-Leibler divergence. The new indicator evaluates how close the actual proportions of groups within a collection are to predefined optimal values. Afterwards, the concept of Coverage is defined, as a function of Relative Balance, the size of the actual collection and its desired size. With the aid of Relative Balance and Coverage, algorithms are defined to allocate or remove accessions to either optimize a collection or a genotyping effort. These definitions and methods are utilized as a proof of concept with the CIMMYT Triticum collection.

The proof of concept

The CIMMYT Triticum accessions and their composition balance:

The CIMMYT (Centro Internacional de Mejoramiento de Maíz y Trigo) Wheat Germplasm Bank conserves 174,553 accessions of bread wheat, durum wheat, triticale and barley from more than 100 countries (https://www.cimmyt.org/tag/germplasm-bank). Currently the collection holds 138,282 accessions of *Triticum* spp.

As a proof of concept for optimization methodology, a simplified classification of the triticum collection is shown in Table 1. Although the groups described are somewhat subjective, they reflect the nature of the collection and will serve to illustrate the use of our model. The respective tree representation is depicted in Fig. 1, whose leaves are the end-groups. The *Triticum* accessions were divided using two criteria that we called Class and Type.

Class refers to the cultivation/breeding status of the accession, divided into Breeders, Cultivar, Landrace and Wild categories. The Breeders class included those accessions classified by the genebank key STATDESC as “Breeders", “Mutant", “Segregant" and “Genetic Stock". The Cultivar class included those accessions classified as “Improved cultivar". The Landrace class included those classified as “Traditional landrace". The Wild

class comprised those classified as “Wild", “Weedy" and “CWR" (crop wild relative).

Type was defined rather as a taxonomic and use-oriented criterion. It was divided into “Bread”, “Durum”, “Other” and “All” categories. The Bread type included the cultivated accessions belonging to the *Triticum aestivum* species. The Durum type comprised the cultivated accessions belonging to the *Triticum turgidum* subsp. *durum* group. Those classified as “Other” comprised cultivated forms that could not be identified as either Bread or Durum wheat, such as *T. monococcum* and *T. timopheevii*. The “All” type belongs to all wild species of *Triticum* in the wheat bank.

Our methods were applied to the simplified information described in Table 1, for “Class” and “Type”, with totals, actual and optimal percentages for the CIMMYT Wheat Germplasm Bank. The Kullback-Leibler divergence (relative entropy) calculated with Equation 2 was 0.08, a number with information units called bits. This number corresponds to a relative balance RB, calculated with Equation 3, of 0.988. This indicates that the collection composition is very close to the defined optimal state.

The size of the actual collection is 135,236 accessions with complete information for classification (Table 1). Let us suppose that there is the possibility to increase the collection to 150,000 individuals because there are more resources. Then application of Equation 3 for coverage (C) gives a value of 0.89, which can be increased by extending the collection size and improving the balance. For optimization, let us assume that the collection size will be extended by 5000 accessions. The question is how to allocate those accessions in the end groups, or, in other words, what sort of accessions and in what amount would be received by the genebank. By using the optimization strategy described in Materials and Methods, and with the use of R implementation, a suggested allocation would be: Breeders-Other 417; Cultivar-Other 520; Wild-All 4063. With these allocations, the relative balance will increase from 0.988 to 0.993.

Now, let us consider a reverse scenario where 5000 accessions must be removed from the collection because of lack of resources. The question is, what end groups must be reduced and in what amounts? By using the optimization process, the suggested removals are distributed as follows: Breeders-Durum -2256; Cultivar-Bread -1609; Cultivar-Durum -726; Landrace-Durum -409. The relative balance will increase from 0.988 to 0.991.

Finally, suppose that we desire to reach a perfectly balanced collection, with 150,000 accessions, and we have the freedom to remove or add any number to each end group. Deciding the number of additions or removals from each end group is a rather arithmetic process, and the result is as follows, where negative numbers mean accession removal: Breeders-Bread 6566; Breeders-Durum -38780; Breeders-Other 945; Cultivar-Bread -6772; Cultivar-Durum -1242; Cultivar-Other 1048; Landrace-Bread 9717; Landrace-Durum -1442; Landrace-Other -107; Wild-All 9340.

One question regarding the optimization of a collection composition would be how to deal with the allocation of accessions when we have not defined optimal proportions for certain end groups. A possible strategy could be to keep the actual proportions on those branches. However, a series of factors would influence the decisions on what sort of accessions to incorporate, such as availability, demand by breeders and genetic redundancy within groups.

Genotyping optimization in a large tree

The most complete tree in the genebank after *Triticum* includes *xTriticosecale*, *Aegilops*, *Agropyron*, *Leymus*, *Aegilotriticum*, *Haynaldia*, *Amblyopyrum*, and the *Tritordeum genus*. The tree was pruned to have at least 10 individuals per leaf with a total of 217 end-groups ending at different levels of classification, because pruned leaves belong to undefined groups. A small number of end-groups with less than 10 accessions were retained because they mainly represent important species.

An objective of the MasAgro-Biodiversidad Project (https://www.cimmyt.org/projects/masagro-biodiversidad/) is to genotype the whole genebank. However, only a limited number of accessions of more than 170,000 present in the genebank can be genotyped. So far, 54,690 accessions have been genotyped by the DArTseq technology, and 10,000 will be genotyped next year. To optimize resources having good representativeness of the genotyped accessions, we want to have approximately the same proportion of genotyped accessions as the proportion of accessions in each end-group. This is a proportional sampling based on the size of the groups, with optimization of the relative balance and the coverage. Therefore, the proportions of accessions in the end groups were considered as the ideal or optimal ones for the application of the formulas and algorithms. The goal is to decide how many more accessions should be genotyped within each end-group.

A table representation of the tree with 217 end-groups (Table S1) shows the number of accessions per end-group (N), and the currently genotyped number of accessions. The grouping criteria were: genus, species, subspecies, breeding status and geographic origin.

The genus criterion included: *Aegilops*, *Aegilotriticum*, *Agropyron*, *Ambly-opyrum*, *Haynaldia*, *Leymus*, *Triticum*, *Tritordeum*, *xTriticosecale*, and Undetermined. The tree included 32 levels for species and 18 levels for subspecies. The breeding status comprised of 17 categories in the STATDESC criterion of the genebank, including "Breeders line", "Genetic stock" and "Improved cultivar", among others. The geographic areas were classified into America (AMERICA), Asia excluding China (ASIA), Balkans and Eastern Europe (BEUROPE), China (CHINA), Central and West Asia and North Africa (CWANA), Oceania countries excluding America (OCEANIA), Western Europe (WEUROPE), and the remaining world countries (OTHER). The list of countries defined for each geographic area can be found in the Supplementary Material.

With the actual number of genotyped accessions, the divergence is 0.361, the relative balance is 0.979 and the coverage is 0.828 (assuming an ideal target size of 64,690 accessions, i.e. the actual one of 54,690 plus 10,000). Although the first genotyped accessions were selected without clear criteria, the relative balance and coverage are fairly good. The new 10,000 accessions that will be genotyped this year can be allocated into the end-groups by maximizing the relative balance using the proposed methodology.

Table 1. Composition of the Triticum collection with two criteria for 135,236 accessions. The Breeders class comprises those accession with the criterion STATDESC classified in the passport data as either “breeders", “mutant", “segregating" or “genetic stock". The Cultivar class comprises all accessions classified as “cultivar". The Landrace class includes all accessions classified as “landrace", whereas the Wild class includes those accessions classified as “wild", “weedy" and ”CWR" (crop wild relatives). The levels of Type include Bread, Durum and Other for cultivated materials, whereas for wild materials it comprises *T. aestivum, T. boeoticum, T. monococcum, T. timopheevi, T. turgidum, T. urartu*, hybrid material and unclassified species (*T.* spp.).

Class	Type	Total	Actual %	Optimal %
Breeders	Bread	30934	22.87	25
Breeders	Durum	9289	6.87	4
Breeders	Other	555	0.41	1
Cultivar	Bread	36772	27.19	20
Cultivar	Durum	4242	3.14	2
Cultivar	Other	452	0.33	1
Landrace	Bread	38283	28.31	32
Landrace	Durum	7442	5.50	4
Landrace	Other	1607	1.19	1
Wild	All	5660	4.19	10

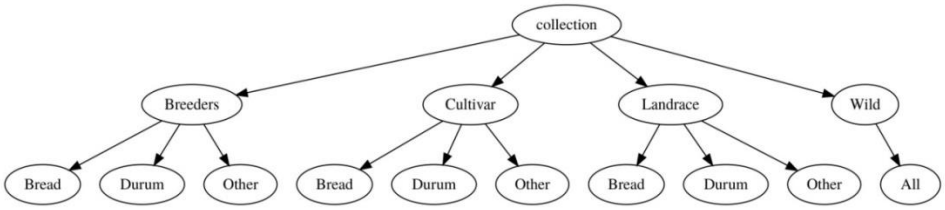

Fig 1. Hierarchical representation of the basic crop composition of Triticum species in Table 1.

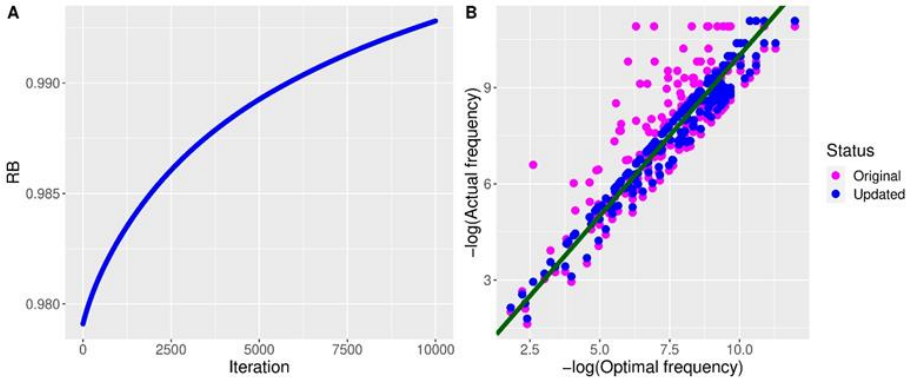

Fig 2. Optimizing the choice of end-groups for genotyping 10,000 accessions: **A** monotonic increase in relative balance (*RB*) through the optimization algorithm along 10,000 accession allocations, **B** relationship between optimal and actual genotyped end-group frequencies before (Original) and after (Updated) the optimal choice of 10,000 accession allocations to be genotyped within end-groups. The straight line represents the ideal end-group frequencies in the genotyped acessions.

By using the function *allocateAccessions* () in the R script, we get the number of allocations for genotyping in each end-group (Supplementary Material Table 1). In brief, it is necessary to increase the number of genotyped accessions in 102 of 217 end-groups with 1 to 3337 accessions with a mean of 98 accessions and a median of 9 accessions. The relative balance after increasing the genotyped accessions by 10,000 is 0.993, while the divergence and the coverage are 0.124 and 0.993, respectively. Compared with the original values, the divergence was reduced to one third from the original value, while the relative balance and coverage were increased by 1.4%.

Fig 2A shows the monotonic increasing of Relative Balance through the optimization process. Fig 2B represents the actual vs optimal frequencies of accessions within end groups before (Original) and after genotyping 10,000 accessions (Updated) following the methodology proposed in the paper. The solid green line represents the ideal placement of the points under

optimal genotyping frequencies. The new proportion of genotyped individuals in the end groups are close to the green line of equality with the optimal distribution. This happens for almost all end-groups more or less uniformly.

Discussion

The indexes proposed in this work are mathematically consistent, with a firm theoretical underpinning based on information theory. Furthermore, they are incorporated into algorithms for collection composition, implemented in R scripts. Nonetheless, the proposed formulas require the definition of optimal proportions for the different groups stored in a genebank. De Oliveira Silva et al. (2019) highlighted the need to also consider biological, environmental and cultural conditions as unavoidable logistic issues.

Thus far, there are few approaches to objectively determine those proportions, with reliance on subjective assessments by bank curators and specialists. In an extended review of next-generation sequencing and its probable effect on genebank management and utilization, Van Treuren and Van Hintum (2014) concluded that the revolutionary advances in the field of sequencing will affect the nature of its services to the user community. But it is not only NGS that will affect the future of genebanks. Given the advent of further improvements in molecular technologies, genebank managers will need to consider the complete genome as well as gene composition and gene expression.

The representativeness of accession genotyping from a genebank can be assessed and optimized following the methods herein described. If the goal is to represent the proportions of end-groups in a genebank in the process of genotyping, then there is no need to define optimal proportions, because the actual ones in the genebank can be used as ideal proportions in the process of optimization. This would not preclude setting genotyping goals that are not oriented to represent the genebank, but to direct the effort towards certain basic or applied research objectives.

We expect that quantitative techniques will be implemented in the short term, to make the optimal proportions as objective as possible. As a proof of concept, the CIMMYT Wheat Germplasm Bank collection was analyzed. Although only two criteria were considered, Class and Type, it provides an example of a well-balanced collection considering the optimal percentages defined by experts, and the relative balance, which is 0.988, and close to optimal setting. Inspection of actual and optimal percentages (Table 1), indicates that the composition of the genebank could be improved by adding more CWR (crop wild relative) accessions. Addition of crop wild relatives would improve the collection balance and would increase the potential for contributing to the never-ending requirements of breeding for disease resistance and tolerance to abiotic stress.

The methods herein proposed were used to allocate genotyping goals in the end-groups of the collection that includes *Triticum* plus eight related genera. The optimized allocations were targeted in 102 of the 217 end-groups. This systematic optimization will facilitate decisions regarding what should be genotyped and will help to make more rational and efficient use of genotyping resources. In general terms, the main idea is to avoid under- or over-representation of certain groups.

Another potential application of the methods herein proposed, is to assess the representation of a genebank in a core subset, in terms of how well it mirrors the group proportions of the genebank.

An issue that has been discussed, but not aimed to be resolved in this paper, is the definition of groups and their optimal percentages in a genebank. For genebanks to be compared in terms of Relative Balance, a consensus must be reached in both definitions. Van Treuren et al. (2009) propose a set of experts and stakeholders to define the optimum collection, and it looks like the best option. Nowadays, more information is generated for the accessions, i.e., molecular and environmental data, and then it has to be considered. Group definition is an aspect to be approached depending on the crop and objectives of the genebank, being either welfare, productivism, or preservationist.

Materials and Methods

Optimizing the composition and genomic characterization of a collection

Crop composition index

The first formula to calculate the imbalance in a crop collection was proposed by Van Treuren et al. (2009). This collection composition imbalance index (*CCII*) is defined as:

$$CCII = \sum_{i=1}^{n} \frac{|A_i - O_i|}{A+O}, \quad (1)$$

where A_i and O_i are the actual and ideal (optimal) numbers of accessions in the *i-th* end group, *A* is the total number of accessions in the collection, and *O* is the desired total. End groups are the end leaves of the diversity tree (Van Treuren et al., 2009) and can be defined at different levels of hierarchy for different objectives. Whereas Equation 1 can be useful in the process of optimizing the composition of a collection, it has two features that require either improvement or re-definition:

(i) When $A_i \leq O_i$ for all *i*, the formula becomes $(\sum_{i=1}^{n} O_i - \sum_{i=1}^{n} A_i)/(A+O)$, which is the same as $(O-A)/(A+O)$. Under this scenario, the value of *CCII* will be invariant, regardless of the proportions of accessions assigned to each end group. Thus, it only measures departure of the actual size of the collection from a specified target size.

(ii) The second feature does not have the same relevance as the first one. However, it reduces the consistency of the parameter. When $A = O$, *CCII* may range from 0 (no imbalance) to 1 (maximum imbalance) according to Van Treuren et al. (2009). However, *CCII* cannot reach the upper limit of 1. Let us consider one example. Assume a collection with an actual number of 1000 accessions, being equal to the optimal number, i.e., $A = O$. Consider that the optimal proportions for three end groups are 0.85, 0.10, and 0.05. The most extreme situation would be a collection with 100% of its accessions allocated in the third end group, i.e., the group whose optimal proportion is 5%. In this extreme case, the *CCII* value will be 0.95, and the maximum of 1 cannot be reached.

In our research, we propose an improved index, which has consistent mathematical properties, inherited from the Kullback-Leibler divergence, a metric related to information theory. Besides its help in measuring the composition balance

of a collection, it can be applied to monitor and optimize genomic characterization.

Balance and coverage of a genebank collection

Let a and o be the actual and ideal proportions for the end groups, represented by the leaves of a tree as a hierarchical depiction of a crop collection, with $o_i > 0$ for all *i*. Define the Kullback-Leibler divergence of the ensemble of actual proportions from the optimal ones as follows:

$$D = \sum_{i=1}^{n} a_i log_2 \frac{a_i}{o_i}, \quad (2)$$

Where log_2 is the base 2 logarithm. Being a measure of divergence, Equation 2 measures the imbalance of the proportions of end-groups in a collection, as compared to the optimal proportions. The minimum value attained by *D* is 0, when the actual proportions equal the optimal ones. It can be shown that the maximum value attained by *D* (see Supplementary Material) is $-log_2(\min(o))$, where $\min(o)$ is an o_i such that $o_j \leq o_i$ for all *i*. Those limits allow us to define the relative balance (*RB*) of a crop collection as follows:

$$RB = 1 + \frac{D}{log_2(\min(o))}, \quad (3)$$

where $\min(o)$ has already been defined. *RB* ranges from 0 to 1, where 0 indicates that the actual proportions of the end groups have the maximum divergence from the optimal ones, whereas a value of 1 indicates a perfect match between actual and optimal proportions.

If, additional to the optimal proportions, a desired size of the collection is defined, and the number is equal to or larger than the actual size of the collection, a measure of coverage, *C*, can be defined, which combines the information of proportions and collection size:

$$C = RB\left(\frac{A}{O}\right), \quad (4)$$

where *A* and *O* are actual and optimal collection sizes, with $A \leq O$. The value of *C* ranges from 0 to 1, where 0 is the minimum coverage due to the maximum divergence (*D*) and 1 indicates that both, proportions and collection size, equal the optimal ones. While it is not possible to define the threshold to distinguish between a low and high Relative Balance, the well-defined range allows comparison between collection compositions or different scenarios for a given collection.

Extension to assess characterization

The set of equations herein proposed to assess relative balance (Equation 3) and coverage (Equation 4) can be used as well to measure the representativeness of a set of characterized accessions from a genebank. Under this setting, a_i is redefined as the proportion of characterized accessions from the *i-th* end-group, and o is the proportion of accessions in a genebank belonging to the *i-th* end-group. The term *A* is defined as the total number of characterized accessions, and *O* is either the number of accessions in the genebank, or a pre-defined target number for characterization.

The interpretation of *RB* for this application will be how well represented is the end-groups in the set of characterized accessions and *C* will be interpreted as the coverage of the characterization effort.

Optimization

There are several scenarios we can consider as optimization cases in the composition of a genebank collection. One of them would be the case where there is a collection of size *A*, and there are resources to add *K* accessions, to have a final size of *A* + *K*. One could ask in which end groups and in what amounts *K* accessions could be allocated, if there are available biological materials of the different types. In this case, a sequential process can be implemented, where *K* allocations are performed, choosing each allocation under the criterion of maximum increase in relative balance until reaching the *K-th* allocation.

Another scenario is the reverse one, where we need to eliminate *K* accessions, so as to have a collection of size *A* - *K*. In this case, a sequential process with the removal of one accession in each step, with the criterion of maximum *RB* would be performed *K* times.

A third scenario can be described as follows: suppose there is freedom and resources to remove or add any number of accessions for each end group, to reach a complete balance and a desired target size. This process would not require any elaborate analytical tool, since it simply involves adding or removing accessions to equalize the actual numbers to the optimal ones.

Optimization tools for the three scenarios, besides calculation of relative balance and coverage, have been implemented in a publicly available R script: (https://github.com/mathgenome/SeedBank).

Conclusion

Germplasm banks worldwide require optimization efforts for a systematic and efficient use of resources directed towards the preservation and characterization of crop diversity. We propose a measure of adequation of the structure of a germplasm bank to previously set goals in terms of group distribution within a given crop and its relatives, embodied in the concept of Relative Balance. The proposed indicator is based on the Kullback-Leibler divergence and is mathematically consistent to assess how well the group proportions in a crop collection fit to a given representation goal. The notion of Relative Balance is the base of yet another definition, called Coverage, which combines group proportions and genebank size. These indicators, besides their potential use as criteria for the balance of accessions in a bank can be used for optimizing genotyping efforts through allocation of accessions to the sample set to be characterized. Optimization strategies are defined for either allocation or removal of accessions for conservation or genotyping. These strategies plus the indicator calculations are implemented in a publicly available R script. The formulas and strategies proposed herein, were applied as a proof of concept to the CIMMYT wheat collection.

Acknowledgements

This research was supported by the Bill and Melinda Gates Foundation, through the CGIAR Research Program MAIZE. MAIZE receives W1&W2 support from the Governments of Australia, Belgium, Canada, China, France, India, Japan, Korea,

Mexico, Netherlands, New Zealand, Norway, Sweden, Switzerland, U.K., U.S., and the World Bank.

References

Acuña-Matamoros C, Reyes-Valdés MH (2018) Comparison of optimization methods for core subset selection from a large collection of mexican wheat landraces characterized by SNP markers. Plant Genetic Resources: Characterization and Utilization. 16(3):228- 236.

Brown AHD (1989a) Core collections: a practical approach to genetic resources management. Genome. 31(2):818-824.

Brown AHD (1989b) The case for core collections. In: Brown AHD, Frankel DH, Marshall DR, Williams JT (eds) The use of plant genetic resources. Cambridge University Press, Cambridge, pp 136-156.

De Oliveira Silva R, Ahmadi BV, Hiemstra SJ, Moran D (2019) Optimizing ex situ genetic resource collections for european livestock conservation. Journal of Animal Breeding and Genetics. 136(1):63-73.

Engels J, Arora R, Guarino L (1995) An introduction to plant germplasm exploration and collecting: planning, methods and procedures, follow-up. In: Guarino L, Ramanatha Rao V, Reid R (eds) Collecting plant genetic diversity. Technical guidelines, CAB International, Wallingford, UK, pp 31-63.

FAO. 2014. Genebank Standards for Plant Genetic Resources for Food and Agriculture. Rome

Frankel OH, Brown AHD (1984) Plant genetic resources today: a critical appraisal. In: Holden JHW, Williams JT (eds) Crop Genetic Resources: Conservation and Evaluation, Allen and Unwin, London, UK, pp 149-257.

Frese L, Palmé A, Kik C (2014). On the sustainable use and conservation of plant genetic resources in Europe. Report from Work Package 5 "Engaging the user Community" of the PGR Secure project "Novel characterization of crop wild relative and landrace resources as a basis for improved crop breeding"

Goodman MM (1990) Genetic and germplasm stocks worth conserving. Journal of Heredity. 81(1):11-16.

Laghetti G, Pignone D, Sonnante G (2008) Statistical approaches to analyze gene bank data using a lentil germplasm collection as a case study. Agriculturae Conspectus Scientificus. 73(3):175-181.

Maxted N, Kell S, Ford-Lloyd B, Dulloo E, Toledo A (2012) Toward the systematic conservation of global crop wild relative diversity. Crop Science. 52(2):774-785.

Radu-Liviu Sumalan, Sorin-Ion Ciulca, Renata-Maria Sumalan, Sorina Popescu (2021) Vegetable Landraces: The "Gene Banks" for Traditional Farmers and Future Breeding Programs [Online First], IntechOpen, DOI: 10.5772/intechopen.96138. Available from: https://www.intechopen.com/online-first/vegetable-landraces-the-gene-banks-for-traditional-farmers-and-future-breeding-programs

Reyes-Valdés MH, Burgueño J, Singh S, Martínez O, Sansaloni CP (2018) An informational view of accession rarity and allele specificity in germplasm banks for management and conservation. PLoS ONE. 13(2):e0193346.

Shands HL (1990) Plant genetic resources conservation: the role of the gene bank in delivering useful genetic materials to the research scientist. Journal of Heredity. 81(1):7-10.

Singh N, Wu S, Raupp WJ, Sehgal S, Arora S, Tiwari V, Vikram P, Singh S, Chhuneja P, Gill BS, Poland J (2019) Efficient curation of genebanks using next generation sequencing reveals substantial duplication of germplasm accessions. Sci Rep. 9:650. https://doi.org/10.1038/s41598-018-37269-0

Smale M, Jamora N (2020) Valuing genebanks. Food Security https://doi.org/10.1007/s12571-020-01034-x

Van Treuren R, Van Hintum T (2014) Next-generation genebanking: plant genetic resources management and utilization in the sequencing era. Plant Genetic Resources. 12(3):298-307.

Van Treuren R, Engels J, Hoekstra R, Van Hintum T (2009) Optimization of the composition of crop collections for ex situ conservation. Plant Genetic Resources. 7(2):185-193.

Vincent H, Wiersema J, Kell S, Fielder H, Dobbie S, Castañeda-Álvarez NP, Guarino L, Eastwood R, León B, Maxted N (2013) A prioritized crop wild relative inventory to help underpin global food security. Biological Conservation. 167:265-275.