Chapter 4

Development of Core Set of Wheat (*Triticum spp.*) Germplasm Conserved in the National Genebank in India

Manoranjan Dutta, B.S. Phogat, Sandeep Kumar, Naresh Kumar, Jyoti Kumari, Avinash C. Pandey, T.P. Singh, R.K. Tyagi, Sherry R. Jacob, Kalyani Srinivasan, I.S. Bisht, M. Karale, Mamata Yadav, Pankaj Sharma, Geeta Kumari, Tariq Aftab, Y.S. Rathi, Amit K. Singh, Sunil Archak, K.V. Bhat, D.C. Bhandari, Y.P.S. Solanki, Dhiraj Singh, and Kailash C. Bansal

Abstract Plant genetic resources, the source of genetic diversity provides a broad genetic foundation for plant breeding and genetic research, however, large germplasm resources are difficult to preserve, evaluate and use. Construction of core and mini core collections is an efficient method for managing genetic resources and undertaking intensive surveys of natural variation, including the phenotyping of complex traits and genotyping of DNA polymorphisms allowing more efficient utilization of genetic resources. A mega characterization and evaluation programme of the entire cultivated gene pool of wheat conserved in the National Genebank, India was undertaken. Wheat accessions with limited seed quantity, were multiplied in the off-season nursery at IARI Regional Station, Wellington during rainy season 2011 and the entire set of 22,469 wheat accessions were characterized and evaluated at CCS HAU, Hisar, Haryana during winter season 2011–12 for 34 characters including 22 highly heritable qualitative, and 12 quantitative parameters. The core sets were developed using PowerCore Software with stepwise approach and grouping method and validated using Shannon-Diversity Index and summary statistics. Based on Shannon-Diversity index, PowerCore with stepwise approach was found better than PowerCore with grouping. The core set included 2,208 accessions comprising 1,770 *T. aestivum*, 386 *T. durum*, and 52 *T. dicoccum* accessions as a representative of the total diversity recorded in the wheat germplasm. The core set developed will

M. Dutta • B.S. Phogat • S. Kumar • N. Kumar • J. Kumari • A.C. Pandey
T.P. Singh • R.K. Tyagi • S.R. Jacob • K. Srinivasan • I.S. Bisht • M. Karale
M. Yadav • P. Sharma • G. Kumari • T. Aftab • Y.S. Rathi • A.K. Singh
S. Archak • K.V. Bhat • D.C. Bhandari • K.C. Bansal (✉)
National Bureau of Plant Genetic Resources, Pusa Campus, New Delhi 110 012, India
e-mail: director@nbgr.ernet.in; kailashbansal@hotmail.com

Y.P.S. Solanki • D. Singh
CCSHAU, Hisar, Haryana, India

be further validated at different agro-climatic conditions and will be utilized for development of mini core set to enhance the utilization by wheat researchers and development of climate resilient improved varieties.

Keywords Characterization • Core collection • Genebank • Germplasm • PowerCore • *Triticum* • Wheat

Introduction

Wheat (*Triticum aestivum* L.) is the most widely cultivated food crop worldwide with an area of 220.39 million hectares and production of 704.08 million tonnes reported during 2011–12 (FAOSTAT 2012). In India, it is the second most important staple food crop after rice, grown in an area of 29.90 million hectares with a total production of 94.88 million tonnes and productivity of 3,140 kg/ha in 2011–12 (DWR Annual Report 2013). The Indo-Gangetic plains comprising the states of Punjab, Haryana, Uttar Pradesh and Rajasthan together account for nearly 85 % of total wheat production in the country. India is probably one of the few countries in the world where three wheat types namely *T. aestivum*, *T. durum* Desf., and *T. dicoccum* Schuebl. are grown although the major area (90 %) is under bread wheat (*T. aestivum*). Bread wheat is grown in all the wheat growing areas while durum wheat is largely grown in Central and Peninsular India mostly under rainfed conditions. In recent years, semidwarf durum wheat varieties have also become popular in Northern India, particularly in Punjab and Haryana. The dicoccum wheat is grown in Maharashtra and Karnataka on an area of about 0.5 million hectares.

Wheat originated in the Fertile Crescent area of south-western Asia among the first domesticated food crops around 8,000 years ago. The north-western end of Indian subcontinent, the fold between Hindukush and Himalaya is regarded as the secondary centre of origin of hexaploid wheat (Vavilov 1926). Archaeological records from many parts of India also revealed cultivation of wheat since the Harappan period (2300–1750 B.C.).

Abundant plant germplasm resources, a rich source of genetic diversity provides a broad genetic foundation for plant breeding and genetic research. However, large germplasm resources are also difficult to preserve, evaluate and use (Holden 1984). Therefore, establishing a core collection (CC) is a favored approach for the efficient exploration and utilization of novel variation in genetic resources (Hodgkin et al. 1995; Zhang et al. 2011). The concept of a CC was first proposed by Frankel (1984) and later developed by Brown (1989a, b). Frankel (1984) defined a core collection as a limited set of accessions representing, maximum diversity with minimum

repetitiveness, the genetic diversity of a crop species and its wild relatives. The core collection could serve as a working collection which could be extensively evaluated. It involves the selection of a subset from the whole germplasm by certain methods in order to capture the maximum genetic diversity of the whole collection while minimizing accessions and redundancy. Frankel and Brown (1984) and Brown (1989a, b) developed this proposal further and described methods to select a core subset using information on the origin and characteristics of the accessions. In developing the core collection, the first issue was its size, second, the grouping of accessions in the entire collection and third, the number of accessions to be selected from a group and fourth the sampling theory. Brown (1989a) using sampling theory of selectively neutral alleles, argued that the entries in a core subset should be ~10 % of the total collection with a ceiling of 3,000 per species. This level of sampling is effective in retaining 70 % of the alleles of the entire collection. The hierarchy of grouping begins with the classification suggested by taxonomy (species, subspecies, and races) followed by assigning accessions to major geographic groups (country, state), climate, or agro-ecological regions. The clustering within the broad geographic group could be done to sort accessions into clusters. A germplasm collection with abundant discriminating data would require multivariate clustering to form groups of similar accessions (Zeuli and Qualset 1993). The number of accessions selected from each class will depend on the sampling strategy used. A good core set should capture maximum genetic diversity with a minimal number of genotypically redundant entries and should be small. Brown (1989a) proposed three procedures based on groups sizes, constant (C), proportional (P) and logarithmic strategies (L). Subsequently, Franco et al. (2005, 2006) proposed that efficiency of sampling for allocation of accessions to different groups could be improved by using diversity- dependent (G) strategy. Of the four strategies, strategy G was reported superior to P strategy (Hodgkin et al. 1999; Yonezawa et al. 1995). Since the original concept of Frankel (1984), core collections have been established in many crop species.

The National Genebank of India currently conserves 31,007 accessions of wheat germplasm comprising 19,116 indigenous and 11,891 accessions of exotic origin. However, the available diversity has not been adequately evaluated and extensively used in wheat improvement due to the large size of germplasm collection. Proper evaluation is feasible only for the traits which can be scored easily and do not show genotype by environment (G x E) interactions. Recognizing this, the present study was aimed to develop the core collection of cultivated wheat germplasm conserved in the National Genebank (NGB) based on characterization and preliminary evaluation data at one representative site with a view to reduce the genebank collection to a manageable level for facilitating utilization of germplasm in applied research.

Material and Methods

Experimental Site and Material

The experiment was conducted during winter season 2011–12 at CCS HAU, Hisar, located at 29°–10' N latitude, 75°–46' E longitude and an elevation of 215.2 m asl. The soils were sandy loam having pH range of 7.5–8.0. The study material included entire gene bank holding of cultivated wheat accessions. For the purpose of core set development, bread wheat (*T. aestivum*), durum wheat (*T. durum*), and emmer wheat (*T. dicoccum*) were grown for agronomic characterization. A set of 22, 663 accession of wheat were grown in Augmented Block Design (Federer 1956) with 8 checks representing different species viz, C 306, PBW 343, DBW 17, RAJ 3765, DWR 1006, UAS 415, DDK 1025, and DDK 1029. The checks were replicated in each of the 114 blocks of 200 accessions each. Each accession was grown in three rows of 2 m length and plant to plant spacing of 25 cm. Standard agronomic practices were followed to raise a healthy crop.

Traits Studied

All the accessions were characterized for 34 important traits, 22 qualitative and 12 quantitative, as outlined by NBPGR minimal descriptors and complete set of observation were recorded for 22,469 accessions. The qualitative characters included early growth vigour (EGV), growth habit (GH), flag leaf angle (FLA), foliage colour (FC), waxiness on leaf blade (WLB), waxiness on leaf sheath (WLS), waxiness on peduncle (WP), waxiness on spike (WS), glume pubescence (GP), auricle colour (AC), auricle pubescence (AP), awnedness (WA), awn length (AL), awn colour (AC), glume colour (GC), spike shape (SS), spike colour (SC), spike density (SD), grain colour (GC), grain shape (GS), grain texture (GT) and grain width (GW). The quantitative traits included, days to 75 % spike emergence (SE), days to 90 % maturity (DM), plant height (PH), effective tillers per plant (EF_T), spike length (SL), number of spikelets per spike (SLS), no. of grains per spike (GRS), grain weight per spike (GRW), 1,000 grain weight (TGRW), dry matter yield per m row length (DMY), grain yield of 1 m row length (GY) and harvest index (HI).

Statistical Analysis

The “PowerCore” (<http://genebank.rda.go.kr/powercore/>) software developed by the Rural Development Administration (RDA), South Korea, that uses the advanced M (maximum) strategy with a heuristic search for establishing core sets possessing

the power to represent all alleles or classes, was used in the present study. It creates subsets representing all alleles or observation classes, with the least allelic redundancy, and ensures a highly reproducible list of entries. This approach has recently been used in developing core set from large rice and foxtail millet collection (Chung et al. 2009; Gowda et al. 2013). It effectively simplifies the generation process of a core set while significantly cutting down the number of core entries, maintaining 100 % of the diversity as categorical variables. Core collections are considered to represent the genetic diversity of the initial collection if the following two criteria are met: (1) no more than 20 % of the traits have different means (significant at $\alpha=0.05$) between the core collection and the entire collection and (2) Coincidence Rate (CR) is retained by the core collection in no less than 80 % of the traits (Hu et al. 2000). The design, concept and implementation strategy of “PowerCore” and the validation on the outcome in comparison with other methods have been well described by Kim et al. (2007). PowerCore by default classifies the continuous variables into different categories based on Sturges rule (Sturges 1926), which is described as: $K = 1 + \log_2 n$, where n = number of observed accessions. However, the software also allows modification of this rule to make desired number of classes for the continuous variables. Once classification of the continuous variables is performed, the software takes into account all classes, without omission of any of its variables. It thus, possesses the capability to cover all the distribution ranges of each class.

Results and Discussion

Genebank Material

Characterization of 22,469 wheat accessions revealed skewed distribution for certain qualitative as well as quantitative characters. Among the qualitative traits the gene bank accessions were skewed for absence of glume pubescence, presence of awns, straw coloured awns, white glume colour and tapering spike shape. Among the quantitative characters, the skewness was observed for traits such as grain length (GL) and grain width (GW) that exhibited highly biased distributions.

Core Set Development

Many approaches for selecting core collections have been proposed and used e.g. M-Strat (Gouesnard et al. 2001), Genetic distance sampling (Jansen and Van Hintum 2007), Power Core (Kim et al. 2007) and Core Hunter (Thachuk et al. 2009). Similarly core has been developed using several kinds of data ranging from

Table 4.1 Species wise description of the core collections developed by different approaches out of entire wheat collection

Species	No. of accessions			
	Entire	Core-P	Core-G	Core-M
<i>T. aestivum</i>	18,101	64	1,215	1,770
<i>T. durum</i>	3,871	53	489	386
<i>T. dicoccum</i>	497	31	209	52
Total	22,469	148	1,914	2,208

Core-P core developed by Powercore default approach, *Core-PG* core developed by Powercore with grouping approach, *Core-PM* core developed by Powercore with modified stepwise approach

genealogical data in the Czech spring wheat (Stehno et al. 2006), agronomic data in groundnut (Upadhyaya 2003; Upadhyaya et al. 2003) and molecular data or integration of data in bread wheat (Balfourier et al. 2007) and in rice (Borba et al. 2009; Yan et al. 2007). PowerCore is a new and a faster approach for developing core collection, which effectively simplifies the generation process of a core set with reduced number of core entries but maintaining high percent of diversity compared to other methods used. In this study, core set was developed with agronomic traits using power core with some modifications. The PowerCore could produce only 64 accessions out of entire wheat accessions (22,469) with default programme without any manual classification and forced selection of entry into the core. Therefore, a modified strategy was followed to make around 8–10 % of entire collection including maximum diversity and minimum redundancy. The method was stepwise random selection using PowerCore with cut-off fixed at around 10 %. With this strategy the core set of 2,208 accessions was developed comprising 1,770 *T. aestivum*, 386 *T. durum*, and 52 *T. dicoccum* accessions (Table 4.1).

Evaluation of Core

Evaluation of core was done by comparing with the other approach, classification and grouping of wheat accessions based on passport data and geographical information (stratified random sampling). The accessions without passport data were classified by hierarchical method of clustering using Euclidean distance and Ward's clustering method. Subsequently all the groups were analysed using PowerCore and then the selected accessions were merged to make the core collection. PowerCore successfully selected 1,914 accessions of the entire wheat germplasm. This consisted 1,215, 489, and 209 accessions of *T. aestivum*, *T. durum*, and *T. dicoccum*, respectively (Table 4.1).

Validation of Core

The core sets developed by three strategies [i.e. species-specific PowerCore (Core P), modified PowerCore (Core PM) and PowerCore involving stratified random sampling based on passport and clustering (Core PG)] were validated by different criteria based on summary of statistics. Means of the entire collection and core subset were compared using Newman-Keul's procedure (Newman 1939; Keuls 1952) for the 12 traits. The homogeneity of variances of the entire collection and core subset was tested with the Levene's test (Levene 1960). It is worth noting that the HCC method gave the same range, minimum and maximum values for the core set generated and the entire collection, indicating its capability to capture almost all of the existing variations. In order to compare the efficiency of "PowerCore" for developing core collection with modified stepwise method and PowerCore with grouping approach method, mean and statistical parameters for entire population, core developed using "PowerCore" and core developed using PowerCore with modified strategy of stepwise method and PowerCore with grouping were compared. The results showed that there was no significant difference ($\alpha=0.05$) for the means of all traits between core and entire collections. The variances of the entire collection and core subset were homogeneous only for five traits viz. days to maturity, plant height, grains per spike, biomass and harvest index. The reason might be due to the large number of germplasm in the entire collection in comparison to that of the core collection. The range of the characters was the same in the entire collection as well as in the core collection implying that the core captured extreme diversity of the total collection (Table 4.2). Four statistical parameters viz., MD (%), VD (%), CR (%) and VR (%), were analyzed using "PowerCore" to compare the mean and variance ratio between core and entire collections. The percentage of the significant difference between the core sets and the entire collection was calculated for the mean difference percentage (MD%) and the variance difference percentage (VD%) of traits. Coincidence rate (CR%) and variable range (VR%) were estimated to evaluate the properties of the core set against the entire collection (Hu et al. 2000).

Mean Difference Percentage (MD %) – which is estimated as:

$$MD(\%) = \frac{1}{m} \sum_{j=1}^m \frac{Me - Mc}{Mc} \times 100$$

Where, Me=Mean of entire collection; Mc=Mean of core collection, and m=number of traits.

Variance Difference (VD %) – estimated as:

$$VD(\%) = \frac{1}{m} \sum_{j=1}^m \frac{Ve - Vc}{Vc} \times 100$$

Where, Ve=Variance of entire collection, Vc=Variance of core collection, and m=number of traits.

Table 4.2 Descriptive statistics for quantitative traits and their validation in entire and core collection using PowerCore-M approach

Traits	Mean		Sig*	Variance		F value	Sig**	Std. Dev.		Minimum		Maximum	
	Entire	Core-PM		Entire	Core-PM			Entire	Core-PM	Entire	Core-PM	Entire	Core-PM
SE	109.8	111.3	ns	130.8	267.6	8.77	s	11.44	16.36	54	54	164	164.00
DM	146.5	147.4	ns	81.66	141.4	5.87	ns	9.04	11.89	120	120	199	199
PH	114.9	115.8	ns	517.7	740.4	0	ns	22.75	27.21	33.5	33.5	197.8	197.8
ET	13.79	14.83	ns	14.42	37.26	70.45	s	3.8	6.1	3	3	48.4	48.4
SL	11.22	11.48	ns	4.21	8.2	11.03	s	2.05	2.86	2.72	2.72	24.6	24.6
SS	19.79	20.25	ns	5.65	14.76	51.82	s	2.38	3.84	5.4	5.4	60.4	60.4
GS	45.56	45.45	ns	131.6	276.7	0.78	ns	11.47	16.63	3.8	3.8	111.8	111.8
GW	1.79	1.88	ns	0.37	0.85	19.93	s	0.61	0.92	0.04	0.04	7.16	7.16
TW	39.81	41.93	ns	151.3	355.3	32.42	s	12.3	18.85	2.72	2.72	204.1	204.1
DM	353.4	347.4	ns	11,120	20,376	6.08	ns	105.5	142.7	20	20	1,042	1,042
SY	91.06	88.71	ns	1,463	2,578	9.8	s	38.26	50.77	1.33	1.33	283.2	283.2
HI	23.92	24.44	ns	93.37	156.8	1.76	ns	9.66	12.52	0.81	1.17	58.34	58.15

Core-PM core developed by Powercore with modified stepwise approach

*Significant at 5 % level, **Significant at 1 % level

Coincidence rate (CR %) – estimated as:

$$CR(\%) = \frac{1}{m} \sum_{j=1}^m \frac{Rc}{Re} \times 100$$

Where, Re=Range of entire collection, Rc=Range of core collection, and m=number of traits.

CR% indicates whether the distribution ranges of each variable in the core set are well represented.

Variable rate of CV (VR %) – estimated as:

$$VR(\%) = \frac{1}{m} \sum_{j=1}^m \frac{CVc}{CVe} \times 100$$

Where, CV_e=Coefficient of variation of entire collection, CV_c=Coefficient of variation of core collection, and m=number of traits.

VR% allows a comparison between the coefficient of variation values existing in the core collections and the entire collections, and determines how well it is being represented in the core sets.

Hu et al. (2000) reported that an MD% smaller than 20 %, in his case 10.07 %, effectively represented the entire collection. The high value obtained for coincidence rate (CR) percentage (95.57 %) suggests that the core attained using the HCC method could be adopted as a representative of the whole collection. In this case, the estimated value for MD% was –6.25, which indicated that there is no difference in the mean values of entire and core collections. VD% was estimated to be 49.04, indicating that the variance for the entire and the core populations are not the same. The CR% obtained was 96.06 which suggests that the core has captured all accessions from all the classes and, thus, is a representative of the entire collection. High VR% (53.87) indicated that the coefficient of variation in the core set is higher compared to entire collections for all the variables. The coefficient of variance in core developed using PowerCore was highest in the case of PowerCore with grouping followed by PowerCore with modified approach and entire collection for all the descriptors. The histogram comparing CV for the entire and core sets is shown in Fig. 4.1. High value obtained for CR% (96.06) suggests that the core obtained using the heuristic approach method could be adopted as a representative of the whole collection.

Shannon-Weaver Diversity Index

The descriptor and descriptor states are parallel to the locus and alleles, respectively, in morphological evaluation. Allelic evenness and allelic richness are the most commonly used parameters for measuring diversity. The allelic evenness in

Fig. 4.1 Coefficient of variation (%) in entire, modified core (Core-PM) and group based core collection (Core-PG) for different traits. *DSE* days to 75 % spike emergence, *DM* days to 90 % maturity, *PH* plant height, *EFT* effective tillers per plant, *SL* spike length, *SLS* spikelets per spike, *GRS* grains per spike, *GRW* grain weight per spike, *TGW* 1,000 grain weight, *DMY* dry matter yield per m row length, *SY* seed yield per m row length and *HI* harvest index

this study was measured using the Shannon–Weaver diversity index, whereas the allelic richness was measured by counting the descriptor states for each descriptor without considering their individual frequencies. The Shannon–Weaver diversity index (H') was computed using the phenotypic frequencies to assess the phenotypic diversity for each character.

$$H = -\sum_{i=1}^n p_i \cdot \ln p_i$$

where p_i is the proportion of accessions in the i th class of an n -class character and n is the number of phenotypic classes for a character. A comparison of Shannon–Weaver (Shannon and Weaver 1949) diversity index for the entire collection, core developed using PowerCore, core developed using modified power core with step-wise approach and PowerCore with clustering method also indicated a high diversity for all the quantitative traits in core developed using PowerCore-M compared to core developed using PowerCore-G approach, except for a few variables, where it was observed at par (Fig. 4.2).

Fig. 4.2 Validation of modified core (Core-PM) and group based core collection (Core-PG) in comparison to entire collection by Shannon diversity index for quantitative traits (traits same as given in Fig. 4.1)

Conclusions

PowerCore is a new and faster approach for developing core collection, which effectively simplifies the generation process of a core set with reduced number of core entries while maintaining high percent of diversity compared to other methods used. Using PowerCore as a tool, three sets of core collections viz. Core P, Core PM and Core PG have been developed. Due to its high Shannon-diversity index, Core PM proved to be the best. These core sets can be further grown with involvement of breeders to select the genotypes with desired background suiting to their requirement. The core sets can be used as a guide for developing trait specific reference/core sets and subsequent allele mining. The best core set could be used as an initial starting material for large-scale genetic base broadening. Thus, it can be concluded that this modified heuristic algorithm can be applied for the selection of genotype data (allelic richness), the reduction of redundancy and the development of approaches for more extensive analysis in the management and utilization of large collection of plant genetic resources.

Acknowledgments We acknowledge with thanks the financial support from the National Initiative on Climate Resilient Agriculture (NICRA) project of the Indian Council of Agricultural Research (ICAR). The technical guidance received from Bioversity International, South Asia Office, New Delhi in the use of software is gratefully acknowledged.

Open Access This chapter is distributed under the terms of the Creative Commons Attribution Noncommercial License, which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

- Annual Report (2013) Directorate of Wheat Research, Karnal, Indian Council of Agricultural Research, India
- Balfourier F et al (2007) A worldwide bread wheat core collection arrayed in a 384-well plate. *Theor Appl Genet* 114:1265–1275
- Borba TCO et al (2009) Microsatellite marker-mediated analysis of the EMBRAPA rice core collection genetic diversity. *Genetica* 137:293–304
- Brown AHD (1989a) Core collections: a practical approach to genetic resources management. *Genome* 31:818–824
- Brown AHD (1989b) The case for core collections. In: Brown AHD et al (eds) *The use of plant genetic resources*. Cambridge University Press, Cambridge, pp 136–155
- Chung HK et al (2009) Development of a core set from a large rice collection using a modified heuristic algorithm to retain maximum diversity. *J Integr Plant Biol* 51:1116–1125
- FAOSTAT (2012) Food and Agricultural Statistics. faostat.fao.org
- Federer WT (1956) Augmented (or hoonuiaku) design. *Hawaii Plant Rec* 2:191–208
- Franco J et al (2005) A sampling strategy for conserving genetic diversity when forming core subsets. *Crop Sci* 45:1035–1044
- Franco J et al (2006) Sampling strategy for conserving maize diversity when forming core subsets using genetic markers. *Crop Sci* 46:854–864
- Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber WK et al (eds) *Genetic manipulation: impact on man and society*. Cambridge University Press, Cambridge, pp 161–170
- Frankel OH, Brown AHD (1984) Current plant genetic re-sources-a critical appraisal. In: Chopra VL et al (eds) *Genetics: new frontiers, vol IV*. Oxford & IBH Publ, New Delhi, pp 1–13
- Gouesnard B, Bataillon TM et al (2001) MSTRAT: an algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. *J Hered* 92:93–94
- Gowda J et al (2013) Use of heuristic approach for the development of a core set from large germplasm collection of foxtail millet (*Setaria italica* L.). *Indian J Plant Genet Res* 26:13–18
- Hodgkin T et al (1995) Core collections of plant genetic resources. John Wiley & Sons and Co-Publishers IPGRI and Sayce Publishing, Chichester, UK
- Hodgkin T et al (1999) Developing sesame core collections in China and India. In: Johnson RC, Hodgkin T (eds) *Core collections for today and tomorrow*. CSA/IPGRI, Rome, Italy, pp 74–81
- Holden JHW (1984) The second ten years. In: Holden JHW, Williams JT (eds) *Crop genetic resources conservation and evaluation*. George Allen and Unwin Publication, London, pp 277–285
- Hu J, Zhu J, Xu HM (2000) Methods of constructing core collections by stepwise clustering with three sampling strategies based on the genotypic values of crops. *Theor Appl Genet* 101:264–268
- Jansen J, van Hintum T (2007) Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce. *Theor Appl Genet* 114:421–8
- Keul SM (1952) The use of the “studentized range” in connection with an analysis of variances. *Euphytica* 1:112–122
- Kim KW et al (2007) Power core: a program applying the advanced M strategy with a heuristic search for establishing core sets. *Bioinformatics* 23:2155–2162
- Levene H (1960) Robust tests for equality of variances. In: Olkin I (ed) *Contributions to probability and statistics: essays in honour of Harold Hotelling*. Stanford University Press, Stanford, pp 278–292
- Newman D (1939) The distribution of range in samples from a normal population expressed in terms of an independent estimate of standard deviation. *Biometrika* 31:20–30
- Shannon CE, Weaver W (1949) *The mathematical theory of communication*. University of Illinois Press, Urbana

- Stehno Z et al (2006) Genealogical analysis in the Czech spring wheat collection and its use for the creation of core collection. *Czech J Genet Plant Breed* 42:117–125
- Sturges H (1926) The choice of a class-interval. *J Am Stat Assoc* 21:65–66
- Thachuk C et al (2009) BMC core hunter: an algorithm for sampling genetic resources based on multiple genetic measures. *Bioinformatics* 6:243
- Upadhyaya HD (2003) Phenotypic diversity in groundnut (*Arachis hypogaea* L) core collection assessed by morphological and agronomical evaluations. *Genet Resour Crop Evol* 50:539–550
- Upadhyaya HD, Ortiz R, Bramel PJ (2003) Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. *Genet Resour Crop Evol* 50:139–148
- Vavilov NI (1926) Studies on the origin of cultivated plants. *Bull Appl Bot Pl Breed* 14:1–245
- Yan W et al (2007) Development and evaluation of a core subset of the USDA rice germplasm collection. *Crop Sci* 47:869–878
- Yonezawa K, Nomura T, Morishima H (1995) Sampling strategies for use in stratified germplasm collections. In: Hodgkin T et al (eds) *Core collections of plant genetic resources*. John Wiley & Sons and Co-Publishers IPGRI and Sayce Publishing, Chichester, UK, pp 35–53
- Zeuli PLS, Qualset CO (1993) Evaluation of 5 strategies for obtaining a core subset from a large genetic resource collection of durum wheat. *Theor Appl Genet* 87:295–304
- Zhang H et al (2011) A core collection and mini core collection of *Oryza sativa* L in China. *Theor Appl Genet* 122:49–61