Development of Core Germplasm

PowerCore: Many gene banks globally contain untapped resources of distinct alleles which will remain hidden unless efforts are initiated to screen these alleles for their potential use and function. The deployment of useful diversity analysis using core collections has been an area of much interests for researches especially those working in the field of allele mining. The prerequisite of any core collection established is that it captures the complete diversity of the entire collection. The recent advancements in technological tools related to genomics and bioinformatics have made it possible to discover new alleles for any gene of interest. These new techniques also create a further challenge of linking traditional phenotypic information to a larger quantity of sequential and genetic information and to complement activities carried out for germplasm enhancement. Allele mining provides the avenue for the validation of specific gene(s) responsible for a particular trait and mining of the most favorable alleles. A computational tool named "PowerCore" provides support to develop a core set by reducing the redundancy of useful alleles and thus enhancing their richness.

Data preparation

Before the PowerCore is executed, the data set has to be inputted into an Excel spreadsheet.

Data Format

- (i) The first row in general contains the information of variable/character names (e.g. %Accession, VG, BP, BC, LA (Note: A percentage (%) character is placed before the title of the Identification column of accessions to represent each of the accessions in the collection.
- (ii) The symbol ~ when placed before the identity of a variable represents a continuous/quantitative data type (e.g. height).
- (iii) The symbol ~ when placed before the identity of an accession indicates a referential selection, wherein the user decides to retain these accessions in the core set without being validated using the PowerCore.
- (iv) The PowerCore program allows any type of character for data input color can be represented as YELLOW or 'A' or 'a' or a numeric data.
- Note: PowerCore supports blank data but does not incorporate these into the final calculation sample data:

%No	VG	BP	BC	LA	FLA	LC	CC	AuriC	CA
1	1	3	2	2	4	1	1	1	2
2	1	3	3	2	4	1	1	1	2
3	1	3	1	1	2	1	1	1	1
4	1	3	3	2	4	1	1		1
5	1	3	5	1	1	3	1	1	2
6	1	3	3	1	1	1	1	1	2
7	1	3	2	2	3	1	1	1	2
8	1	3	4	2	3	3	3	2	3
9	1	3	2	1	1	1	1	1	1
10	2	3	2	1	1	1	1	1	2
11	1	3	1	2	1	1	1	1	2

The Excel spreadsheet can be copied directly into the interface of the program.

Running the PowerCore

1. To run the installed program, go to the 'START' toolbar, and search for the PowerCore program and click 'Open'.



fig1: Opening the PowerCore from Start Menu

2. Once the program is executed, the following window appears on the screen. Using the mouse pointer, right click on the screen.

PowerCare 1.0				[] [] X	PowerCore 1.0					
fumber or Accessions	Number of Vanidales	Que II	POLY BLANDER TO MOON		Number of Accessions	Number of Variables	🔞 Holp 🔳 s	ingant de Steps IR Steps		
			nghatili at Resolver				And Constant Constant And	n gift did on the unwer		
		Construction of the Science of Science of the Science of the						grindszal Bohechvology, RDA, R. Korea All rights reserv		
Mater a th	a Vianustrate	The feature from the second second second	The second stress store and the second store store and the	BOB OF BOA TOP	📲 start 🛛 🥫 😋 🛢	2 C Pagetralisseter	PowerCore User Ha	Heroroft Escal - Phese	Document - Microsof	

Fig 2: Data import from Excel sheet

- 3. Clear and Paste functions are used when the existing information is replaced with a new data set.
- 4. Copy all function is used for exporting the existing data set to a Clipboard for Excel spreadsheet.
- 5. Append function is used when new information is added to an existing file.

Note: PowerCore accepts various and has no limit for data input size. Data input is based on the resources are available in the user's computer, and not according to the limit of the excel spreadsheet.



Fig 3: PowerCore window after data import.

6. Place the mouse pointer on the top right corner of the window and click Step 1 The crucial step 1 would be converting the quantitative data into classes and to validate the reliability of the data set (e.g. deleting missing/blank data). This is important as in general a continuous data set has no variables and is expressed in real numbers or in integer format.

7. Click 'Classifying' to create classes of each variable determined by the criterion of Sturge's rule. This will allow each accession to be allocated to these created classes. Figure 4 displays the output in the form of a histogram:



Fig 4: Histogram showing classes of variable

8. Click 'Run' to perform the heuristic search. By checking the 'Random' button, the search is performed using the random method – Accessions are selected randomly instead of being selected by the heuristic evaluation function.



Fig 5. RUN window for Heuristic searching

The following figure (Figure 6) shows the steps whereby the heuristic algorithm searches for the best possible accessions to be selected for the core set.

E PowerCore 1.0				
		🗖 Random 🧲 Run	Input II. Step1 IR Step2	
Entry 8 2920 666 5	Number of Accessions 1000 Number of Variables 39 Number of Entries Non-Incuristic Search 52 Max Possible Entries 100 PowerCore 4			
	Efficiency Index Unfilled Diversity Cells 102 Random Branching 0			
	MD% CR% VD% VR%	Click to detail informations		
	Diversity Index			
	V CoreSet V EntireSet			
	PIC			
	C-Count			
Searching	10 %			
🦺 start 🤐	🤨 🕙 🦈 🧮 PowerCore_User_Ma	a Dower core data MI PowerCore 1.0) 🐏 Documenk1 - Microsof 🦷 🍟 untitled - Paint	🛃 😌 🛍 😒 🗊 🐨 😁 - 5:12 PM
Ι				

Fig 6. Heuristic searching for best possible accessions.

The left most panel on the screen are the selected entries (accessions ID as per data set) using the heuristic search. By right clicking the 'Entry' tab, the list could be copied to a clipboard.

The panel displayed shows each variable in the form of a histogram. By right clicking the histogram, a separate table indicating the number of accession for each class, core count and the entire count is displayed.

hit PowerCore 1.0						💶 🖻 🔜
		🗆 Random 🧉 Run	Input 💷 s	itep1 🔛 Step2		
Entry 8 2920 666 5	Number of Accessions 1000 Number of Variables 39 Number of Entries Non-heuristic Search 52 Max Possible Entries 166 Provertine 45	~LL		0		
65 3188 2759 13 1129 2654	Efficiency Index 0.86 Unfilled Diversity Cells 0 Random Branching 0 MD% 6.98 CR% 96.85 VD% 71 VR% 200.51	3 	1309	16.45 19.82	28.55 26.55 91	36 40 37 54
9 2 3158 2325 10 3090 3031 1128 3146	Crowerstity Index Option CroreSet C EntireSet PIC C-Count E-Count	Class Core Count Fr 3.00 ~ 6.36 4 49 6.36 ~ 9.73 6 10 9.73 ~ 13.09 10 45 10.97 10.06 7 24 16.45 ~ 19.62 6 25 16 50 23.18 ~ 26.55 3 4 55 3 4	ntre count 000 188 13 2			
7560 2396 193 2646 2321 76 2607 1682		29.91 × 33.27 0 0 33.27 × 95.64 1 1 36.64 × 40.00 1 1				
716 1587 2421 764 2189 12 744 2897						
1581 678 411 2932 675 2932 675	L S D ** The PowerCore_User_Pla	🍋 power core data 🛛 🖬	PowerCore 1.0	umenti - Microsof 🏾 👸 unbibled - Pa	ant	酒 ● 数 2 単 ● ● ● 513m

Fig 7. Complete accession level detail of core set generated via PowerCore

Click the 'diversity index' tab to display the diversity index using Nei and Shannon & Weaver calculation (Figure 8)

PIC- Nei DI

C Count – Core Set by Heuristic Method

E Count – Entire collection

ht PowerCore	1.0														
			□ R.	andom	E Run				II 📕	nout II. St	tep1 🗟	Step2			
Entry	<u> </u>	Number of Accessions 1000	~SH							0.279					-
2		Number of Variables 39													
3158		Number of Entries													
2325		Non-heuristic Search 52													
10		ProverCore 45													
3090		remarcine is					-				_	_			
3031		Efficiency Index 0.86			Cleb to d	all all last a	and the second								
1128		Unfilled Diversity Cells 0	1000		Pener to u	ecanino	macional			-			10	63	7.0
3146		Random Branching 0	50	4	16		0	99	8	13	13	526	10	100	8 -
773		MD% 6.98 CR% 96.85		01	0		64	600	63	-3	10	ar.	-	on	
1560		VD% 71 VR% 200.51												1	
2396				Character	C-Sh.W.	C-Nei	C-Alele	E-Sh.W.	E-Mei	F-Allele					- Ala
193		Diversity Index		MG	0.842	0.428	*1	0.658	0.358						
2646		Option	1	RD	0.287	0.126	3	0.03	0.01	3					
2321		I✓ CoreSet I✓ EntreSet		DC.	1 102	0.635		0.601	0.970	6					
76		PIC 0.99		DI GC	0.202	0.167	4	0.026	0.01	4					
2607		PAC 0.99		1.0	0.5	0.22	2	0.197	0.000	7					
1682		C-Count 45		EL A	1.042	O EOE	4	0.002	0.000	4					
716		E-Count 1000		I.C.	0.05	0.000	2	0.902	0.023	3					
1587				00	0.307	0.100	3	0.030	0.014	3					
2421				aunter .	0.100	0.120	3	0.03	0.022	3					
764				AURIC	201.02	0.065	2	0.046	0.022	~					
2189				LC.	0.090	0.547	3	0.724	0.505	3					
12				CC CC	0.35	0.105	3	0.051	0.010	3					
744				LS OF	0.888	0.555	3	0.679	0.423	3					
2897				60	0.617	0.618	3	0.66	0.539	3					
1591				56	1.001	0.207		0.086	0.032	4					
678				PE	1.031	0.510	2	0.694	0.445	0					
411				AP	1.061	0.55	5	0.582	0.304	2					
2932				Apec	1.372	0./14	9	1.009	0.572	5					
675				SC	0.35	0.105	3	0.127	0.055	3					
3162				LPC	0.57	0.244	5	0.143	0.049	5					
1674				SLC	0.35	0.165	3	0.095	0.034	3					
2488				SUL	0.213	0.086	3	0.016	0.004	3					
1349				SCC	0.943	0.594	5	0.293	0.179	5					
	-			ET	0.496	0.444	2	0.366	0.255	2					-
- etart	105	D C " C Internet Fr.	C Inter	out the I	191 months	in the second second	I BEL	teres B -	1.00	Microsoft F		Denergene			10-54 654

Fig 8. Display of Diversity Index.

Saving the result

Once data generated for the core set is saved, a new excel sheet is generated by the PowerCore. Filtering of the core set from the entire collection is done, and the core set is automatically marked 'Y' by the software.



📕 Microsoft Exe	cel - TestCore								
Eile Edit	⊻iew Insert F	ormat <u>T</u> ools	<u>D</u> ata <u>₩</u> ir	ndow <u>H</u> elp	2				
0 💕 🖬 🔒		9 📖 🐰 🖻	1 🙇 • 🥥	1 - (≌ - 🤶 X	- A↓ A↓	1 🛄 🛷 📀	🚆 🗄 Arial	• 10 • B <i>I</i>
A1		ore	0.0		10.			100 B	D. D. D.
A	B C	D	E	F	G	Н	1	J K	L M
1 Core 🗣	6No 🔽VG	■BP (BC 	BLSC	🗖 LA	FLA	LC (🖌 CC 💽 AuriC	TCA TIC T
Sort Ascending	1	3	2	1	2	4	1	1 1	2 2
Sort Descending	1	3	3	1	2	4	1	1 1	2 2
(All)	1	3	1	1	1	2	1	1 1	1 1 1
(Top 10)	1	3	3	1	2	4	1	1	1 1)
(Custom)	1	3	5	4	1	1	3	9 9	2 9
Y (Diamina)	1	3	3	1	1	1	1	1 1	5 5 1
(NonBlanks)	1	3	5	1	5	3	1	1 1	5 1 1
9 Y 19	1	5	Ā	5	5	5	5	5 5	3 4
10 Y	1	5	5	1	Ĩ	1	1	5 5	1 1 1
11 V 1	n 5	5	5	1	1	1	1	9 9	5 5 1
12 1	1 1	5	1	1	5	1	1	1 1	5 5
13 V 1	2 4	5	5	5	1	1	9	1 1	á á)
14 V 1	3 4	5	5	4	4	4	4	9 9	5 9 7
15	4 5	5	5	5	4	5	5	5 5	4 4 1
16 4	E 1	5	5	5	5	5	5	6 6	4 4 9
10		5	3	5	2	3	5	6 6	
10 1	7 4	5	5	4	-	4	-		
10	0 4	5	5			5	4	A A	5 5 7
19		5	2					N N	5 6 1
20 1	9 1	3	2	1	1	1	1	1 1	2 1

X /	Microsoft	Excel - T	estCore												
:2	Eile Eo	lit <u>V</u> iew	Insert Fg	rmat <u>T</u> ools	Dat	a <u>W</u> indow	<u>H</u> elp								
1	💕 🔒	🔒 🔒 l e	🖻 🛕 🗳	۵ 🔏 ا 🔏 🖻	Â↓	<u>S</u> ort				1 🛍	I 🛷 🔞	🙄 🕴 Ari	ial	- 10	-
_	A1	-	f∡ 'Coi	re		Eilter			•	, I	Auto <u>F</u> ilter				
	A	В	С	D		Subtotals				9	5how All		K	L	
1	Core	%No	VG	BP		Validation					- Advanced Filt	er	AuriC	CA	IC
2		- I - 5	4	ა ნ		Text to Colu	INDE			-	- 1	4	4	5	5
4		2	1	3		10/200 00/201					1	1	1	1	1
5		4	1	3	1	PivotTable a	and PivotChar	t Report			1	1	· ·	1	1
6	γ	5	1	3		Import Exte	rnal <u>D</u> ata		►		3	1	1	2	1
7		6	1	3		Lįst			•		1	1	1	2	2
8		7	1	3		×MI					1	1	1	2	1
9	γ	8	1	3							3	3	2	3	4
10	Y	9	1	3	¥.,	<u>R</u> efresh Dat	a				1	1	1	1	1
11	Y	10	2	3			*				1	1	1	2	2
12		11	1	3	1	1	2				1	1	1	2	1
13	Y	12	1	3	- 5	1	1		1	,	1	1	1	1	1
14	Y	13	4	ა ნ	- 5	4	4		1		1	1	4	2	4
16		14	9	5	- 5		5		2		1	4	1	1	4
17		16	1	5	- 5				1		1	1	1	1	1
18		17	1	3	- 5	1	1	-	* 3	-	1	1	1	1	1
19		18	1	3	5	1	1		1		1	1	1	5	1
20		19	1	3	2	1	1		1		1	1	1	2	1
21		20	1	3	2	1	1		3		1	1	1	1	1
22		21	1	3	1	1	1		2		1	1	1	1	1
23		22	1	3	1	1	1		2		1	1	1	1	1
24		23	1	3	2	1	1		3		1	1	1	2	1
25		48	1	3	2	1	1		2		1	1	1	1	1
26		49	1	3	2	1	1		1		1	1	[1	1	1
27		150 1 54	1	3	2	1	1		1		1	1	1	1	1
28		<u>1</u> 51	[2	3	[2	[1	[1	["	I	ſ	1	1	11	1	[1

Fig 9. Complete accession level detail of core set generated via PowerCore

Statistical terms & formulas :

1. 'Classifying' to create classes of each variable determined by the criterion of Sturge's rule Sturge's rule = 1 + Log2(n), n : the observed number of accessions.

Number of Accessions : Total number of accessions from the existing collection Number of Variables: Represents the number of characters from the data set

- 2. Non-heuristic Search : A search which does not use any heuristic algorithm (Note : Similar to random search, but results are always repetitive as search is performed sequentially)
- 3. Max Possible Entries : It is the worst case scenario, wherein this is the limit for PowerCore to select the maximum number of entries.
- 4. Efficiency Index : Effectiveness of PowerCore in comparison to the non-heuristic search.

PowerCore Max Possible Entries

- Note: A lower value signifies a more effective search
 - 5. Unfilled Diversity Cells: Status during the filling of the diversity index.
 - 6. Random Branching: Selection of nodes randomly during the selection process of same accessions with same values of minimum evaluation functions, indicating the number of its occurrence.

Analysis with Statistical Indicators:

i) Mean difference (MD) (%) $= \frac{1}{m} \sum_{j=1}^{m} \frac{|Me - Mc|}{Mc} \ge 100$

Me: Mean of entire collection, Mc: Mean of core collection. ii) Variance Difference(VD) (%) = $\frac{1}{m} \sum_{j=1}^{m} \frac{|ve - vc|}{vc} X$ 100

Ve: Variance of entire collection, Vc: Variance of core collection

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

Re: Range of entire collection, Rc: Range of core collection

iv) Variable Rate(VR) (%) = $\frac{1}{m} \sum_{j=1}^{m} \frac{cvc}{cve} \times 100$

CVe: coefficient of variation of entire collection, CVc: coefficient of variation of core collection. m: number of traits

References and suggested readings:

Kyu-Won Kim, Hun-Ki Chung, Gyu-Taek Cho, Kyung-Ho Ma, Dorothy Chandrabalan, Jae-Gyun Gwag, Tae-San Kim, Eun-Gi Cho and Yong-Jin Park (2007). PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets *Bioinformatics*. 23 (16): 2155-2162.

PowerCore User Manual