

# Maximizing Biodiversity in Core Subset Selection

Using Stochastic Local Search

Chris Thachuk

International Center for Maize and Wheat Improvement (CIMMYT), Mexico  
CIHR/MSFHR Bioinformatics Training Program

VanBUG, October 2006

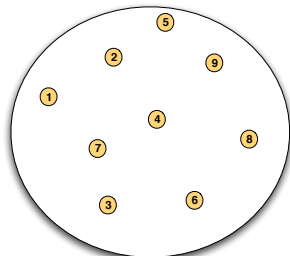


SIMON FRASER  
UNIVERSITY

# Core Subset

## Definition

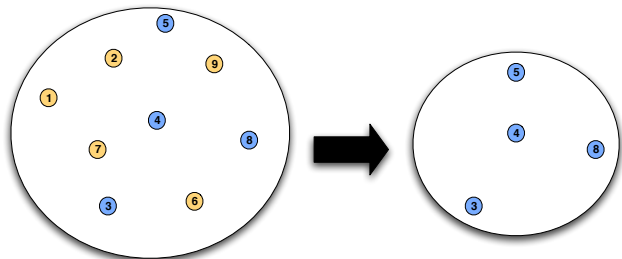
A **Core Subset** is selection of genotypes/phenotypes from a reference dataset, chosen due to their diversity in traits, *genetic distance*, *genetic diversity* or preference of the breeder.



# Core Subset

## Definition

A **Core Subset** is selection of genotypes/phenotypes from a reference dataset, chosen due to their diversity in traits, *genetic distance*, *genetic diversity* or preference of the breeder.



# Why Core Subsets? - Too much data

- Helps Save Money
- Facilitates Genetic Conservation



# Why Core Subsets? - Too much data

- Helps Save Money
- Facilitates Genetic Conservation



# How hard is this Problem?

- For a dataset with 512 genotypes, to choose a of core 10% (51 genotypes), there are:
- $7.3 \times 10^{70}$  possible cores
- It would take the world's fastest computer, many more years to compute than there are grains of sand on Earth's beaches ( $7.5 \times 10^{18}$ )<sup>1</sup>.

---

<sup>1</sup><http://www.hawaii.edi/suremath/jsand.html>



# How hard is this Problem?

- For a dataset with 512 genotypes, to choose a of core 10% (51 genotypes), there are:
- $7.3 \times 10^{70}$  possible cores
- It would take the world's fastest computer, many more years to compute than there are grains of sand on Earth's beaches ( $7.5 \times 10^{18}$ )<sup>1</sup>.

---

<sup>1</sup><http://www.hawaii.edu/suremath/jsand.html>



# How hard is this Problem?

- For a dataset with 512 genotypes, to choose a of core 10% (51 genotypes), there are:
- $7.3 \times 10^{70}$  possible cores
- It would take the world's fastest computer, many more years to compute than there are grains of sand on Earth's beaches ( $7.5 \times 10^{18}$ )<sup>1</sup>.

---

<sup>1</sup><http://www.hawaii.edu/suremath/jsand.html>



# Allele Frequency Matrix

	M1_a1	M1_a2	M1_a3	M2_a1	M2_a2
1	0.5	0.0	0.5	1.0	0.0
2	0.3	0.2	0.5	0.2	0.8
3	0.0	1.0	0.0	0.9	0.1
4	0.1	0.8	0.1	0.4	0.6



# Allele Frequency Matrix

	M1_a1	M1_a2	M1_a3	M2_a1	M2_a2
1	0.5	0.0	0.5	1.0	0.0
2	<b>0.3</b>	<b>0.2</b>	<b>0.5</b>	0.2	0.8
3	0.0	1.0	0.0	0.9	0.1
4	0.1	0.8	0.1	0.4	0.6



# Genetic Distance

- Distance measures quantify the dissimilarity between two genotypes
- May be the measures of choice for breeders



**Warning:** if you don't like equations, don't look right.

$$MR_{xy} = \frac{1}{\sqrt{2L}} \sqrt{\sum_{l=1}^L \sum_{a=1}^{n_l} (p_{xla} - p_{yla})^2} \quad (1)$$

$$CE_{xy} = \sqrt{\frac{1}{L} \sum_{l=1}^L \left(1 - \sum_{a=1}^{n_l} \sqrt{p_{xla} p_{yla}}\right)} \quad (2)$$



# Calculating Genetic Distance

	M1_a1	M1_a2	M1_a3	M2_a1	M2_a2
1	0.5	0.0	0.5	1.0	0.0
2	0.3	0.2	0.5	0.2	0.8

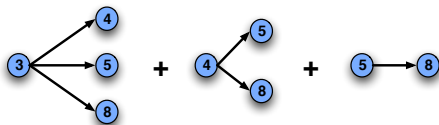
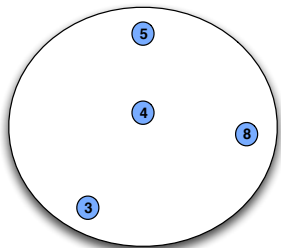


# Calculating Genetic Distance

	M1_a1	M1_a2	M1_a3	M2_a1	M2_a2
1	0.5	0.0	0.5	1.0	0.0
2	0.3	0.2	0.5	0.2	0.8



# Average Genetic Distance of a Population



Take the average



# Genetic Diversity

- Diversity measures look at overall allele richness (not a measure between genotypes)
- May be the measures of choice taxonomists and population geneticists

$$SH = - \sum_{a=1}^A (\hat{p}_a \ln \hat{p}_a) \quad (3)$$



**Another Warning:** if you don't like equations, don't look left.



# Calculating Genetic Diversity

	M1_a1	M1_a2	M1_a3	M2_a1	M2_a2
1	0.5	0.0	0.5	1.0	0.0
[Redacted]					
3	0.0	1.0	0.0	0.9	0.1
4	0.1	0.8	0.1	0.4	0.6



# Calculating Genetic Diversity

	M1_a1	M1_a2	M1_a3	M2_a1	M2_a2
1	0.5	0.0	0.5	1.0	0.0
3	0.0	1.0	0.0	0.9	0.1
4	0.1	0.8	0.1	0.4	0.6



# Existing Strategies

- Current best strategy for Genetic Distance  
**D-Method** (*Franco et al. 2005*)
- Current best strategy for Genetic Diversity  
**MSTRAT** (*Gouesnard et al. 2001*)



# Existing Strategies

- Current best strategy for Genetic Distance  
**D-Method** (*Franco et al. 2005*)
- Current best strategy for Genetic Diversity  
**MSTRAT** (*Gouesnard et al. 2001*)



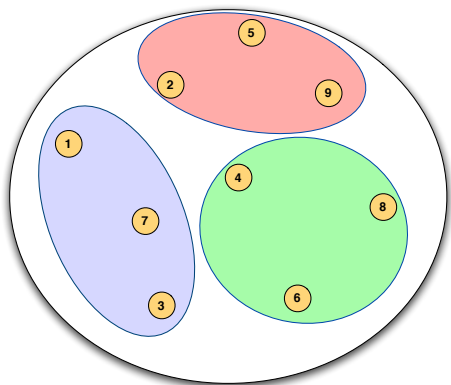
# Existing Strategies

- Current best strategy for Genetic Distance  
**D-Method** (*Franco et al. 2005*)
- Current best strategy for Genetic Diversity  
**MSTRAT** (*Gouesnard et al. 2001*)



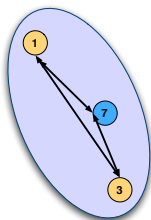
# D-Method Visually

## Step 1 - Stratify dataset into C clusters



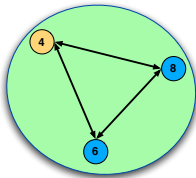
# D-Method Visually

**Step 2,3** - Form core subset by randomly selected accessions from clusters based on proportions of previous step.



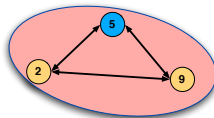
Avg Distance = **a**  
 $P_a = N \cdot a / (a+b+c)$

**1**



Avg Distance = **b**  
 $P_b = N \cdot b / (a+b+c)$

**2**



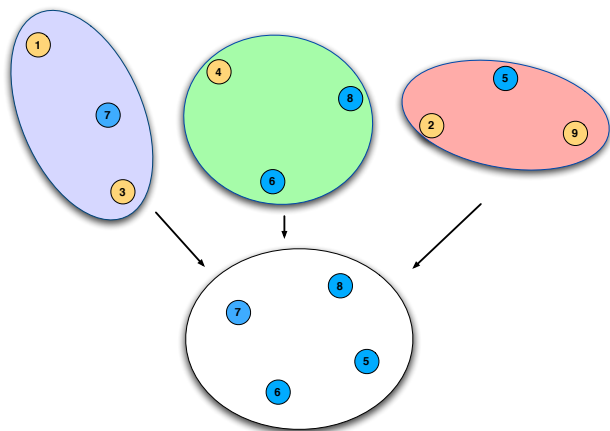
Avg Distance = **c**  
 $P_c = N \cdot c / (a+b+c)$

**1**



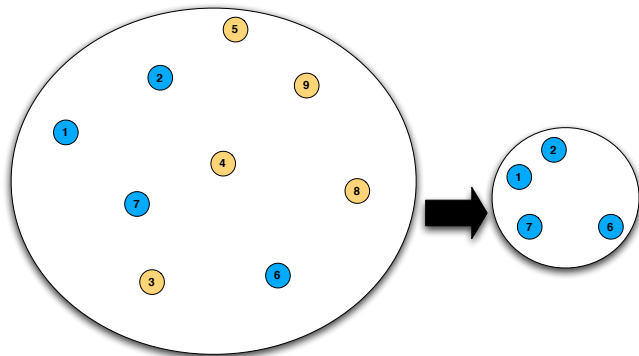
# D-Method Visually

**Step 2,3** - Form core subset by randomly selected accessions from clusters based on proportions of previous step.



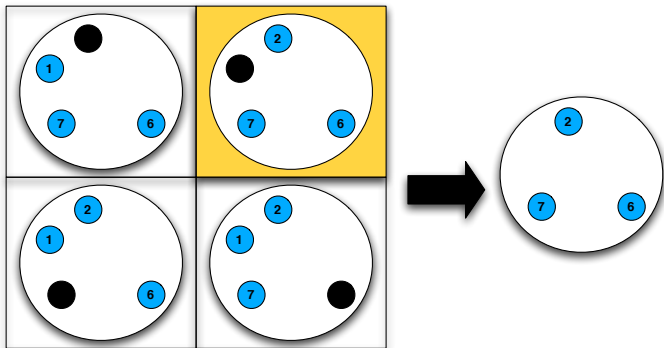
# MSTRAT Visually

**Step 1** - Randomly select N accessions to form initial core



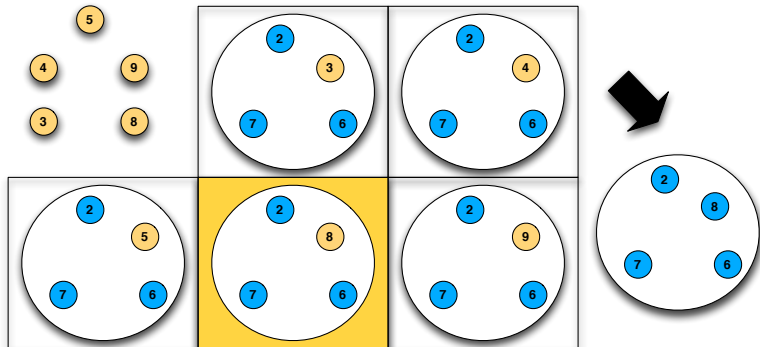
# MSTRAT Visually

**Step 2** - Determine M-score of each subset after throwing out one accession.



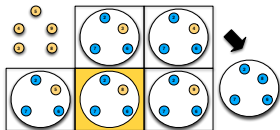
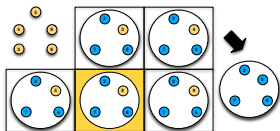
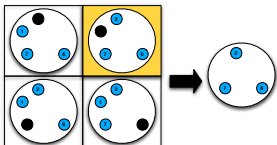
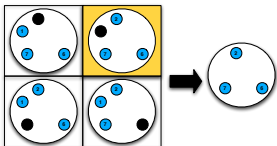
# MSTRAT Visually

**Step 3** - Determine M-score of each subset after adding one accession. Keep the accession which was added to the best subset.



# MSTRAT Visually

**Step 4 - Repeat Steps 2 and 3 until no further improvement.**



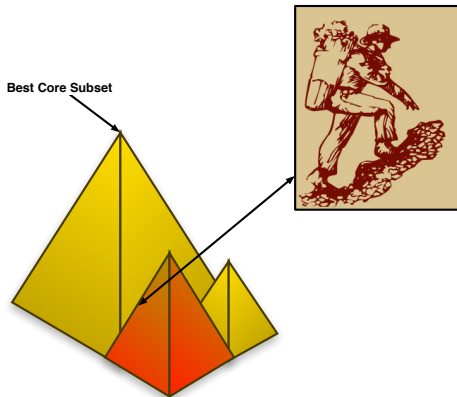
# MSTRAT Weaknesses

- Maximizes only on Diversity
- Uses simple hill climbing algorithm

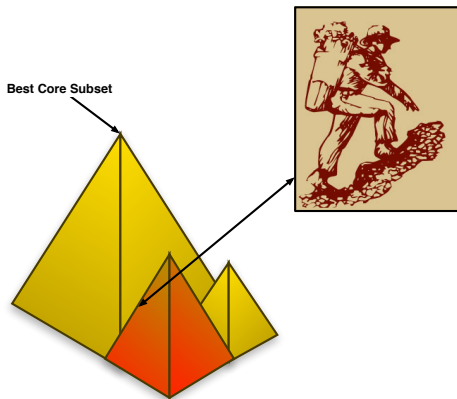


# MSTRAT Weaknesses

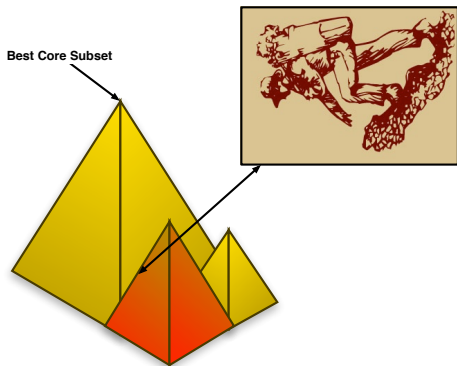
- Maximizes only on Diversity
- Uses simple hill climbing algorithm



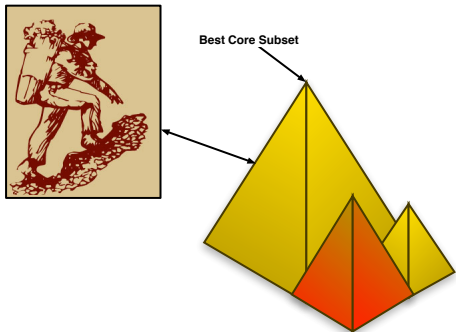
# Idea 1 - More Sophisticated Search



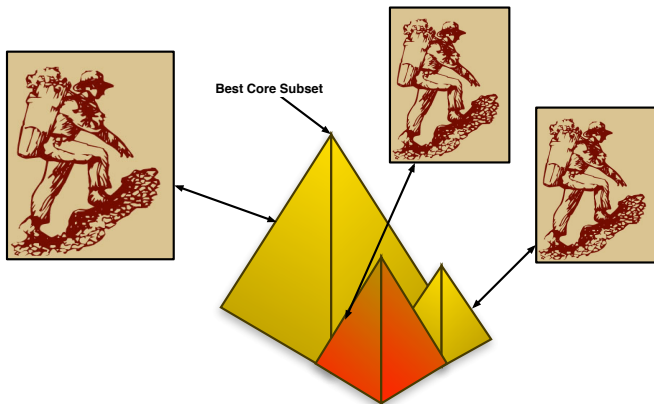
# Idea 1 - More Sophisticated Search



# Idea 1 - More Sophisticated Search



# Idea 1 - More Sophisticated Search



# The Search Method - Replica Exchange Monte Carlo

## Ensemble State

$$C = \{c_1, c_2, \dots, c_M\}$$

## Probability of Replica Exchange

$$\begin{aligned} Pr[C \rightarrow C'] &\equiv Pr[l(c_i) \leftrightarrow l(c_j)] \\ &= \begin{cases} 1 & \Delta \leq 0 \\ e^{-\Delta} & \text{otherwise.} \end{cases} \end{aligned}$$

$$\Delta = (\beta_j - \beta_i)(E(c_i) - E(c_j))$$

Y. Iba, *Extended ensemble monte carlo*, International Journal of Modern Physics C, vol. 12, p. 623, 2001.



# The Search Method - Replica Exchange Monte Carlo

## Ensemble State

$$C = \{c_1, c_2, \dots, c_M\}$$

## Probability of Replica Exchange

$$\begin{aligned} Pr[C \rightarrow C'] &\equiv Pr[l(c_i) \leftrightarrow l(c_j)] \\ &= \begin{cases} 1 & \Delta \leq 0 \\ e^{-\Delta} & \text{otherwise.} \end{cases} \end{aligned}$$

$$\Delta = (\beta_j - \beta_i)(E(c_i) - E(c_j))$$

Y. Iba, *Extended ensemble monte carlo*, International Journal of Modern Physics C, vol. 12, p. 623, 2001.



# Results on Maize Data

Bulk Data Set			
	MR	CE	SH
<b>D-Method</b>	0.503	0.578	4.411
<b>MSTRAT</b>	0.477	0.571	4.493
Local Search	0.572	0.641	4.531



# Results on Maize Data

Bulk Data Set			
	MR	CE	SH
<b>D-Method</b>	0.503	0.578	4.411
<b>MSTRAT</b>	0.477	0.571	4.493
<b>Local Search</b>	<b>0.572</b>	<b>0.641</b>	<b>4.531</b>



# Results on Maize Data

Accession Data Set			
	MR	CE	SH
<b>D-Method</b>	0.653	0.719	4.525
<b>MSTRAT</b>	0.647	0.718	4.579
Local Search	0.695	0.752	4.670



# Results on Maize Data

Accession Data Set			
	MR	CE	SH
<b>D-Method</b>	0.653	0.719	4.525
<b>MSTRAT</b>	0.647	0.718	4.579
<b>Local Search</b>	<b>0.695</b>	<b>0.752</b>	<b>4.670</b>



## Idea 2 - Maximizing Multiple Criteria

<b>D-Method</b>	<b>MSTRAT</b>
Distance	Diversity



## Idea 2 - Maximizing Multiple Criteria

<b>D-Method</b>	<b>MSTRAT</b>
Distance	Diversity



## Idea 2 - Maximizing Multiple Criteria

D-Method	MSTRAT
Distance	+ Diversity

$$F = \alpha CE + (1 - \alpha)SH, \quad 0 \leq \alpha \leq 1$$



## Idea 2 - Maximizing Multiple Criteria

D-Method	MSTRAT
Distance	+ Diversity

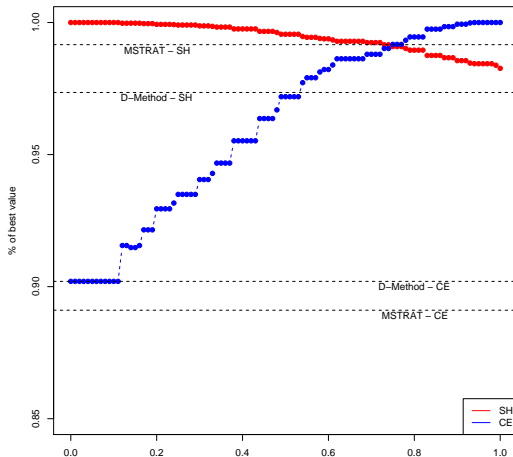
$$F = \alpha CE + (1 - \alpha)SH, \quad 0 \leq \alpha \leq 1$$

$$G = \alpha_1 CE + \alpha_2 MR + \alpha_3 SH + \alpha_4 PN, \quad \alpha_1 + \alpha_2 + \alpha_3 + \alpha_4 = 1$$



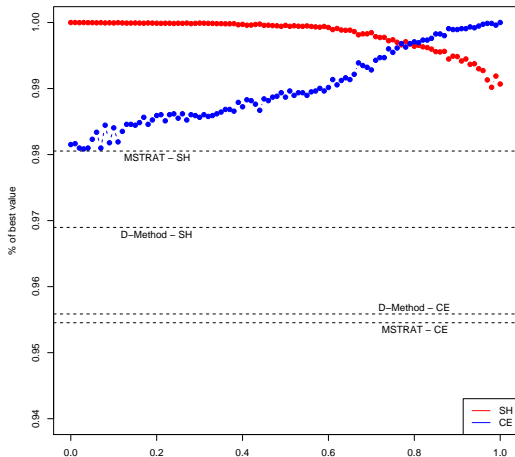
## CE vs SH

## Bulk Dataset

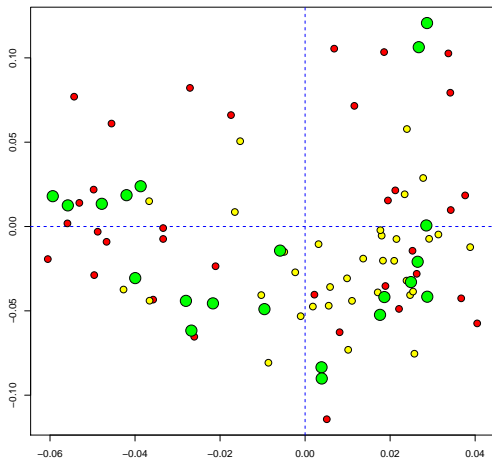


## CE vs SH

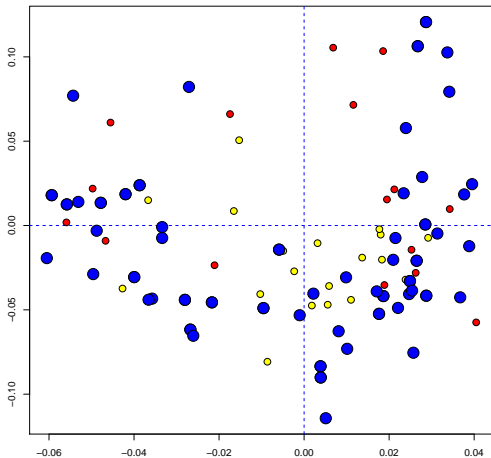
## Accession Dataset



# PCO Analysis - How the cores relate



# PCO Analysis - How the cores relate



# Conclusions

- More sophisticated search techniques enable us to find better cores
- Maximizing Diversity & Distance simultaneously is possible



# Conclusions

- More sophisticated search techniques enable us to find better cores
- Maximizing Diversity & Distance simultaneously is possible



# Future Work

- Maximize cores based on phenotypic traits
- Dynamic core size



## Future Work

- Maximize cores based on phenotypic traits
- Dynamic core size



# Thanks

Guy Davenport

Jose Crossa

Susanne Dreisigacker

International Center for Maize and Wheat Improvement (CIMMYT)

International Rice Research Institute (IRRI)



*CIHR/MSFHR Strategic Training Program in*

**BIOINFORMATICS**



SIMON FRASER  
UNIVERSITY