

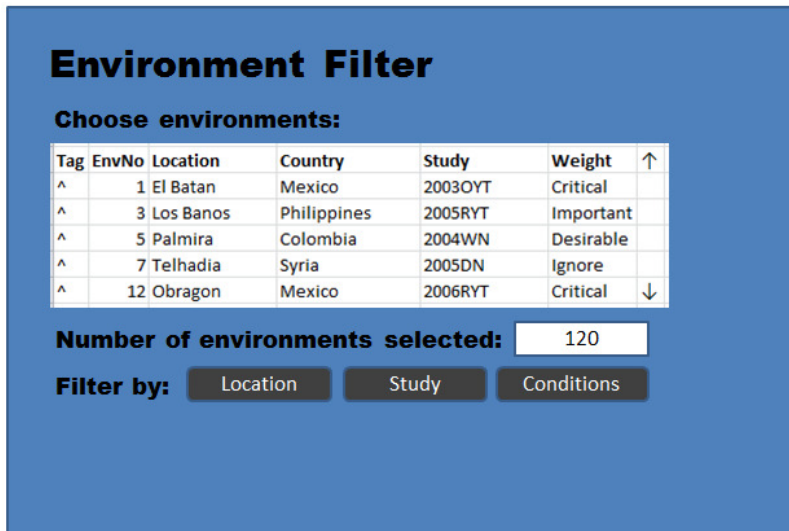
Breeders' cross-study queries for IBDB V2 Phenotyping Data

1. Query for Adapted Germplasm

Find germplasm suitable for specific environmental conditions. (Filter trial environments by the required environmental conditions, and find the germplasm with the best performance of some important trait(s) in those environments).

A. Specify and weight the environments

First we have to get a set $\{E\}$ of *nd_geolocation_ids* for trial environments. The complete list of *nd_geolocation_ids* can be filtered in three ways – on location, on study and on environmental conditions (*nd_geolocationprops*). Start with all environments selected and progressively de-select environments by using any or all of the three filters in any order.



Environment Filter

Choose environments:

Tag	EnvNo	Location	Country	Study	Weight	
^	1	El Batan	Mexico	2003OYT	Critical	↑
^	3	Los Banos	Philippines	2005RYT	Important	
^	5	Palmira	Colombia	2004WN	Desirable	
^	7	Telhadia	Syria	2005DN	Ignore	
^	12	Obragon	Mexico	2006RYT	Critical	↓

Number of environments selected: 120

Filter by:

The list of environments should be in a table viewer so that it can be sorted on any of the columns. The weight is an editable column filled with the weight 'Important' by default for all selected (Tagged) environments and with 'Ignore' for all deselected environments. The user can change the weight of any environment to 'Critical' to give it more weight than 'Important' or to 'Desirable' to give it less weight. Changing the weight of an 'Ignored' environment will select it again. When the final list of environments is used weights W_k , will be assigned as 0,1,2,3 for Ignored, Desirable, Important and Critical environments. The complete list of Environments should be maintained and there should be a switch to show only retained (Tagged) environments or all environments so that excluded environments could be brought back in if necessary.

- Filter on Location: Start with the list of retained environments. If the environment (*nd_geolocation* record) does not have a label (in table *nd_geolocationprop*) with property Location and scale DBID reject it (because we don't know it's location). If it does, show users a list of countries corresponding to environments with numbers of

environments in each country. Allow users to de-select countries with a check box. For any selected country allow users to click on the county to show a list of provinces (first sub-national division) with numbers of environments in each province. Allow users to deselect provinces. Also allow users to click on selected countries and show a list of location names which users can deselect.

Filter Trial Environments on Location

Filter by Location

Tag	Country	No Envs
^	Colombia	2 ↑
^	Mexico	
^	Philippines	
^	Syria	
^	Thailand	
^	Zimbabwe	

Provinces of Philippines

Tag	Province	No Envs
^	Laguna	4 ↑
^	Mindoro Occ	2
^	Mindanao	1
^	Ilocos Norte	1

Environments in Mindoro Occ

Tag	EnvNo	Study
^	3	2005RYT
^	7	2006HB

- Filter on Study: Start with the list of retained environments. Get the list of Studies to which they belong. Build a table view of the Study name, Study title and number of environments in each study. (The table viewer allows users to sort the table on any column). Allow users to deselect studies with a check box. This de-selects all environments from that study.

Filter by Study

Tag/untag Studies to select/deselect all environments in that Study

Tag	Study name	Study title	PI name	IRRIGATION	NoEnvs
^	2003OYT	Irrigated program OYT for 2003	G. Khush	Yes	3
^	2005RYT	Irrigated program RYT for 2005	G. Khush	Yes	6
^	2004WN	Upland program Winter Nursery	G. Atlin	No	1
^	RL00OFT	Rainfed lowland OFT 2000	S. Sakarung	No	16
^	RL01OFT	Rainfed lowland Oft 2001	S. Sakarung	No	15

Add Study conditions

The ‘Add Study Conditions’ button shows users a list of all study conditions in the selected set with a count of the number of studies with that condition (same view as for Environmental Conditions below). These conditions come from the *projectprop* table and include the standard ones like study name, study title, objective, start date, end date, study type and pmkey, but they also include user selected ones like PI, or IP conditions. The user Tags conditions to add to the Study Filter table view.

- Filter on Environmental Conditions: When the user clicks the ‘Conditions’ button on the Environment Filter, build a table view of all Trial Conditions for the selected trial environments from the *nd_geolocationprop* table with a column giving the number of environments which have that label. Tag conditions to add as columns to the Environment Filter table view.

Tag	Condition	Description	NoEnvs
^	TRIAL_LOCATION	Location - selected (name)	10
^	LOCATION_ID	Location - selected (DBID)	8
^	COOPERATOR_ID	COOPERATOR ID -Assigned (DBID)	8
^	COOPERATOR	COOPERATOR NAME	7
^	SEEDING_DATE	Date Seeded	6
o	TRNSPL_DATE	Date Transplanted	6
o	ROWS_PLOT	Number of rows/plot	6
^	ROW_SPACING	Spacing within rows	5
o	PLANT_DENSITY	Plant Density	5
o	NFERT_MGMT	N fertilizer - applied (kg/ha)	4

OK

B. Get the data and compute the environment weights

- Using {E} obtained in part A.
- Get the list of observation units {Q} (*nd_experiment_ids*) evaluated at {E} by filtering *nd_experiment.nd_geolocation_id* on {E}.
- You can get the list of germplasm {G} (*stock_ids*) tested in {E} by linking {Q} back to stocks through the *nd_experiment_stock* linkage table.
- You can get a list of standard variates {T} (*cvterm_ids*) measured at {E} by linking {Q} to phenotype records (through the *nd_experiment_phenotype* linking table) and then linking *phenotype.observable_id* to *cvterm.cvterm_id* you get the list {T} of Standard variable *cvterm_ids* evaluated on germplasm {G} in environments {E}.

You now have a four-way table of ids: {G,T,E,Q} each combination indexing one value relevant to the problem $\{O_{ijkl}: i \text{ in } G, j \text{ in } T, k \text{ in } E \text{ and } l \text{ in } Q\}$ (the l th observation of the j th trait in the k th environment for the i th line).

At this stage the weight classes for the selected environments need to be converted to numerical weights, E_k , adding to 1.0 in the ratios suggested in section A.

C. Set up the Trait Filter. This needs three sections on the same screen

Section 1: Get all values for numeric variates and summarize them for each trait with column headings:

- Trait: Trait name (mouse over description)
- No of Locations in {L} (mouse over for a list of location names)
- No of Lines in {G} observed for that trait any locations in {L}
- No of Observations, since the same line might have been observed several times in a location
- Min: Min observation for numeric variates in {T}
- Median: Median value for numeric variates {T}
- Max: Max value for numeric variates {T}

Trait	No of	No of	No of	Range of Values			Trait Filter		
	Location	Lines	Observations	Min	Median	Max	Condition	Limits	Priority
Protein	3	12	39	3.1	7.3	8.9	between	4,6	Critical
GYLD	7	45	371	1027	2031	3315	>	3000	Critical
Lodging	2	8	20	0	10	80	<	5	Important
HI	6	20	200	0.2	0.5	0.7	>	0.5	Desirable

The trait names should be clickable to pop out the line x location values for each trait in the table viewer (see mockup below).

The ‘condition’ pull down box on the filter should have conditions: ‘drop trait’, ‘keep all’, <, <=, =, >=, >, ‘between’, ‘in’, ‘not in’. The ‘Limits’ box should accept the appropriate number of limits for the condition.

The Priority column should have pull-down boxes with the options Critical, Important, Desirable and Ignore which would affect the weight given to the trait in computing an adaptation index.

Section 2. Get all Character variates and produce a similar display with the ‘Range of values’ columns replaced a text box for each variable containing the distinct observed values for each trait separated by commas.

Most values would be hidden if the list is long so we might have to be able to pop them out or view the list with mouse over.

Trait	No of Locations	No of Lines	No of Observations	Distinct Values Observed	Trait Filter	
					Condition	Limits
Aroma	3	8	24	none, mild basmati, sweet, strong basmati, s	not in	none, sweet
Remarks	7	5	35	stiff straw, open panicles, suffered from aph	drop trait	

The Trait names should be clickable and result in a table view of the Line by Location values. This section would also have a Priority column as in Section 1.

The pop up Line by Location table view would look as follows for a numeric trait:

Line x Location Table for Trait: Protein - Kejall (%)						
Observation	Line			Location		
No	No	GID	Designation	Los Banos	El Batan	Palmira
1	1	9758	IR 71697-59-1-3-1		4.2	
2	2	6885	IR 71697-2-3-2-2	3.7		6.1
3	3	4905	IR 71692-45-2-2-1	4		7.2
4	3	4905	IR 71692-45-2-2-1	3.5	3.7	
5	4	765	IR 71687-48-2-1-1		6	8.9
6	5	8741	IR 71687-11-2-2-1	4		
7	6	6852	IR 71684-36-3-3-2			7.5
8	6	6852	IR 71684-36-3-3-2	6.1	3.3	5.5
9	7	113	IR 71673-55-1-3-1	3.7		6.2
10	7	113	IR 71673-55-1-3-1		3.1	6.3
11	7	113	IR 71673-55-1-3-1	5		7.1
12	8	4530	IR 71456-7-1-2	3.3	5	
13	8	4530	IR 71456-7-1-2	4.1		6.3
14	9	209	IR 71451-2-1-1		3.1	
15	10	876	IR 71449-31-1-2	6.1		4.3
16	10	876	IR 71449-31-1-2	3.7		
17	10	876	IR 71449-31-1-2	4.1	4.2	3.7
18	11	5413	IR 71218-39-3-2	5		
19	11	5413	IR 71218-39-3-2	3.3	3.8	5.7
20	12	2393	IR 71218-5-2-1	4.1	4.2	
21	12	2393	IR 71218-5-2-1		3.9	6.1

Section 3. Get all Categorical Variates and produce a similar display with the ‘Range of values’ columns replaced by two rows for each trait showing:

- Classes: Valid values observed for categorical variates in {T}
- Frequency: Frequency of observed values for categorical variates {T}.

Trait	No of Locations	No of Lines	No of Observations	Observed Values / Frequency					Trait Filter	
				Class 1	Class 2	Class 3	Class 4	Class 5	Condition	Limits
BLB	3	12	39	S 12	MS 18	R 9			>=	MS
EXS	2	8	20	1 3	3 3	5 7	7 6	9 1	in	3,5,7

The Trait names should be clickable and result in a table view of the Line by Location values. This section would also have a Priority column as in Section 1.

There have to be as many Class columns as the number of observed classes for the variable with the highest number of classes observed. The order of the classes is in alphabetic order of class values (even if they are numeric) or in the order provided in the *cvtermprop* table if there is one (so there would have to be an order specified for BLB to get the order shown in the mockup). This order is used in the order conditions.

The observation number is an artificial construction (a record in the table viewer) that simply allows multiple G,T,E combinations. This could occur if there are several replications of the observation in a particular environment. We also need to consider what to do when there is raw data and means data at a site. For the moment I suggest we just show it all and see what makes sense.

The Table viewer has a checkbox column which can indicate records to keep. These would all be checked by default, but could be unchecked for any line where the particular trait score indicated definite rejection irrespective of observations on other traits. We need to keep an overall check list for all the entries to display in the results box. This involves some ‘averaging’ of checks since the Line by Location Tableview shows all observations but we want results for each line. Not sure how to do the averaging, but for a start we can say that rejection on the basis of any observation should reject the line.

D. Display the results

When the user has specified the filters show the result:

- Line: GermplasmID
- Name: Designation
- Trait1: No of observations for each line over all environments
- Trait1: Score for each line for Trait1 over all locations.
- Trait 2

We can define an adaptation score for each line for each trait. Let R_{ijkl} be +1 if observation O_{ijkl} is within limits for trait j, and -1 if it is outside the acceptable limits. Then a score for line i and trait j over all environments k could be $R_{ij..} = \sum_k E_k \sum_l R_{ijkl} / n_{ijk}$ where n_{ijk} is the number of observations of trait j for line l in environment k.

Tag	Line			<u>Protein</u>		<u>GYLD</u>		<u>Lodging</u>		<u>HI</u>		<u>BLB</u>		<u>EXS</u>		<u>Aroma</u>		Combined Score
				Wt=	20	Wt=	30	Wt=	10	Wt=	20	Wt=	10	Wt=	5	Wt=	5	
	No	GID	Designation	NoObs	Score	NoObs	Score	NoObs	Score	NoObs	Score	NoObs	Score	NoObs	Score	NoObs	Score	
^	12	4905	IR 71692-45-2-2-1	8	3.1	27	8.5	0	n/a	27	3.2	12	80	20	15	5	20	12.3
^	17	765	IR 71687-48-2-1-1															11.2
^	3	8741	IR 71687-11-2-2-1															10.4
	15	6852	IR 71684-36-3-3-2															
^	5	113	IR 71673-55-1-3-1															
^	1	5415	IR 71218-39-3-4			32	9.3											
	18	1138	IR 71673-55-1-3-5															
^	8	4530	IR 71456-7-1-2															
	27	4533	IR 71456-7-1-5															
	10	5413	IR 71218-39-3-2															
...																		
	49		IR 71451-2-1-1															
	2		IR 71449-31-1-2													10	18	6.1

We would also like to have a combined score over all traits, but again we have a problem that not all traits are equally important, so in the mockup we have a weight for each trait, T_j . A default could be to have Ignored traits weighted 0, Desirable traits weighted 10, Important traits weighted 20, and Critical traits weighted 40. The Combined Score is just the weighted average of the trait scores for each line across all traits ($R_i \dots = \sum_j T_j R_{ij} / \sum_j T_j$). Then the table could be sorted on any of the columns, but particularly on Combined Score. The user could change the weights in the header of the table and the ranking would change in real time.

We probably want to be able to go back to the Trait Filter and change the settings if we get too many or too few results.

We would want to be able to save lines from this table to a list. This would usually be the top x lines, but the set we want might not always be contiguous so we probably need a check box next to each line (as on the left of the mockup) which can indicate which to save to a list. This check list would inherit the results from viewing individual traits in the Line by location tables.

2. Find Trait Donors

Find germplasm that have acceptable values for a set of specified traits .

A. Specify the set of traits of interest

Browse the trait CV and select the traits of interest. (Chain through the *cvterm_relationships* of type 1225 for terms from cv 1000 or 1040 from cvterm 1001, or for biological traits only, from cvterm 1330. Search jumps to leading matches as you type.)

Find Trait Donors

Specify traits of interest: or browse the ontology:

- ☐ IBDB Class
 - ☐ Crop research Ontology
 - ☐ Crop trait ontology
 - ☐ Agronomic
 - ☐ Grian_yield, Grain yield -dry and weigh (kg/ha)
 - ☐ Plant_height, Plant height - soil to tip at maturity (cm)
 - ☐ Morphological
 - ☐ Biotic stress
 - ☒ BLB, Bacterial leaf blight – visual assessment (score)
 - ☐ Abiotic stress
 - ☐ Grain quality

Selected traits:

- ☒ Grian_yield, Grain yield -dry and weigh (kg/ha)
- ☒ Plant_height, Plant height - soil to tip at maturity (cm)
- ☒ BLB, Bacterial leaf blight – visual assessment (score)

B. Get the Data

For the specified traits (*cvterm_ids* in cv 1040) find all the observations for each trait and get the data (*phenotype* records with matching *observable_id*). Find all the observation units on which each trait is observed (*nd_experiment* records related to the *phenotype* records). Find all the environments where each trait is observed (*nd_geolocation* records with matching *nd_geolocation_ids*). Finally, find all the germplasm for which each trait is evaluated (*stock* records related to *nd_experiment* records).

C. Filter and weight the environments

Filter the environments (starting with the list recovered above) exactly as in Query 1 section A.

D. Specify the acceptable trait values

Set up the trait filter exactly as in Query 1 section C.

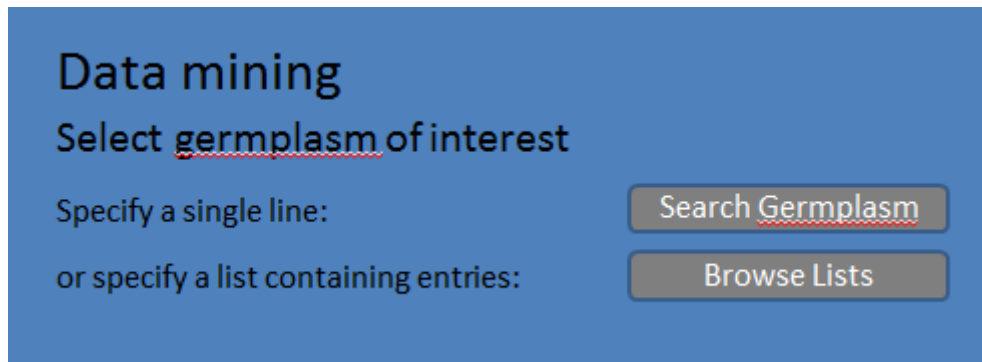
E. Display the results

Display the results and allow the user to save a list of selected germplasm exactly as in Query 1.

3. Data mining

Find all available data for a given list of germplasm. Specify traits of interest, and filter experimental observations according to environments.

A. Specify the germplasm



B. Get the Data and Specify the traits to be reported

The next step is to find all environments where any of the entries are tested, and the list of traits observed in each of those environments. (Map the GIDs from the specified list to the *dbxref_id* field of the stock table, map those stocks to observation units (*nd_experiements*), map the *nd_experiements* to phenotype and get the variables and observations recorded there.)

Produce a table view of traits and the number of lines which have at least one observation of that trait, and the number of environments where they are measured. Select the traits to be reported using the Tag column of the table view.

Tag	Trait		No Lines	No Envs
^	GRAIN_YIELD		10	8
^	PLANT_HEIGHT		10	6
o	MATURITY		8	5
o	FLOWERING		9	3
^	BLB		4	2
o	CHALK		2	1
o	AROMA		1	1

Should have a column for the Trait Description as well.

Specify the reporting traits using the Tag column.

C. Filter the environments

List the environments in the environment selector table viewer together with the number of observations for each trait and for any line in the environment:

Environment Filter

Choose environments:

Tag	Env No	Location	Country	Study	GRAIN_YIELD	PLANT_HEIGHT	BLB	Weight	
^	5	El Batan	Mexico	2005RYT	5	6	2	Important	↑
^	3	Los Banos	Philippine	2004OYT	4	5	4	Important	
^	2	Palmira	Colombia	2002HB	7	8	3	Important	↓

Number of environments selected:

Filter by:

Filter the environments as in Query 1. Section A.

D. Report the results

Obs No	Line	GID	EnvID	Site name	GRAIN_YIELD	PLANT_HEIGHT	BLB
1	IR 71697-59-1-3-1	9758	12	EL Batan	4.3	87	3
2	IR 71697-2-3-2-2	6885	12	EL Batan			2
3	IR 71692-45-2-2-1	4905	12	EL Batan	3.2	76	4
4	IR 71692-45-2-2-1	4905	8	Los banos		89	
5	IR 71687-48-2-1-1	765	8	Los banos	4.3	87	3
6	IR 71687-11-2-2-1	8741	13	Palmira	5.6	102	
7	IR 71684-36-3-3-2	6852	13	Palmira		92	6
8	IR 71684-36-3-3-2	6852	13	Palmira	5.5	88	1
9	IR 71673-55-1-3-1	113	13	Palmira	6.1	103	4
10	IR 71673-55-1-3-1	113					
11	IR 71673-55-1-3-1	113					
12	IR 71456-7-1-2	4530					
13	IR 71456-7-1-2	4530					
14	IR 71451-2-1-1	209					
15	IR 71449-31-1-2	876					
16	IR 71449-31-1-2	876					
17	IR 71449-31-1-2	876					
18	IR 71218-39-3-2	5413					
19	IR 71218-39-3-2	5413			3.3		1
20	IR 71218-5-2-1	2393			4.3	78	2
21	IR 71218-5-2-1	2393	19	Telhadia	5.1	89	3

Allow the user to:

- Add Study Conditions for the studies containing the environemts

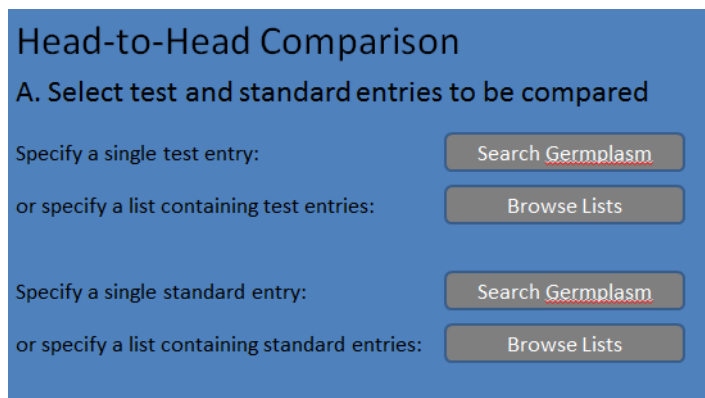
- **Add Environmental conditions for the environments in the table**
- **Save the result to a workbook or csv file.**

4. Head to head comparisons

The head-to-head comparison compares each entry in a list of test entries with each entry in a list of standard entries. Either or both lists might consist of only one entry. The comparison is done pairwise on evaluation data over environments where both entries were evaluated. The key result is the proportion of environments where the test entry surpassed the standard entry for a particular trait.

A. Specify the entries

We need a screen requesting the user to specify a list of (t) test entries and a list of (s) standard entries. These lists can be assumed to have been made in the list manager earlier. However we want to facilitate the case where either or both of the lists consist of a single entry ($t=1$ or $s=1$), and allow the user to specify that entry directly (by a call to search germplasm).



Head-to-Head Comparison

A. Select test and standard entries to be compared

Specify a single test entry:

or specify a list containing test entries:

Specify a single standard entry:

or specify a list containing standard entries:

B. Select the traits and environments for the comparisons

The next step is to find all environments where any pair of entries (test x standard) are tested together (comparative environments), and the list of traits observed in each of those environments.

Produce a table view of traits and the number of the comparative environments where they are measured. Select the traits on which the pairs will be compared using the Tag column of the table view and specify the desirable direction for the comparison of each selected trait.

Tag	Trait	No Envs	Direction
^	GRAIN_YIELD	8	Increasing
^	PLANT_HEIGHT	6	Decreasing
o	MATURITY	5	
o	FLOWERING	3	
^	BLB	2	Decreasing
o	CHALK	1	
o	AROMA	1	

(Note: We have not determined how many pairs are comparable for each trait in any environment, only that the pairs are evaluated for at least one trait in each environment and that the trait was measured on some line in that environment.)

Drop environments where the selected traits were not measured and retrieve the data for each selected trait and each line in each environment. T_{ijk} is the observation of trait i for test line j in environment k if it is measured (else null). S_{ilk} is the observation of trait i for standard line l in environment k if it is measured (else null). For categorical variates let S and T be the order number of the observed class as defined by the alphabetical order of the classes or by the order property of the classes if there is one. (We need to consider the possibility that there may be more than one observation for any trait-line-environment combination. If so we could average these to get a single observation).

Count the number of pairs that are comparable for each environment and each trait (For each i and k , count the number of pairs T_{ijk}, S_{ilk} $j=1,2..t$, $l=1,2...s$ where both elements are not null). Present a table view of environments:

Environment Filter

Choose environments:

Tag	Env No	Location	Country	Study	GRAIN_YIELD	PLANT_HEIGHT	BLB	Weight	
^	5	El Batan	Mexico	2005RYT	5	6	2	Important	↑
^	3	Los Banos	Philippine	2004OYT	4	5	4	Important	
^	2	Palmira	Colombia	2002HB	7	8	3	Important	↓

Number of environments selected:

Filter by: Location Study Conditions

Allow users to select environments as in section A of query 1. The weights can be used to compute a weighted trait difference for each pair, but if you leave it at default a straight average

is computed. (The weight classes are converted to numerical weights E_k as described in Query 1. Section B. Suppose there are m remaining environments.

C. Compute and present the results

For each trait $\{i\}$ and each pair $\{j,l: j=1,2,\dots,t, l=1,2,\dots,s\}$ compute N_{ijl} , M_{ijl} , D_{ijl} and P_{ijl} where N_{ijl} is the number of selected environments where $\{T_{ijk}, S_{ilk}: k=1,2,\dots,m, \text{ are both not null}\}$, M_{ijl} is the number of those environments where $\{T_{ijk} \geq S_{ilk}: k=1,2,\dots,m\}$ if the desirable direction of trait i is increasing or $\{T_{ijk} \leq S_{ilk}: k=1,2,\dots,m\}$ if it is decreasing, $D_{ijl} = r * \sum_k E_k * (T_{ijk} - S_{ilk})$ where $r=+1$ if the desirable direction of trait i is increasing or -1 if it is decreasing and E_k is $1.0/N_{ijl}$ if all selected environments have equal weight or it is $W_k / \sum_k W_k$ if the environments have different weights W_k as described in Query 1 section B. P_{ijl} is the probability that M_{ijl} or more successes occur in a Binomial ($N_{ijl}, 0.5$) population.

<http://www.stat.berkeley.edu/~stark/Java/Html/ProbCalc.htm>

Present the results:

Test Entry	Standard Entry	GRAIN_YIELD						PLANT_HEIGHT						BLB					
		No	No	Mean	Mean	Mean	Pval	No	No	Mean	Mean	Mean	Pval	No	No	Median	Median	Mean	Pval
		Env	Sup	Test	Std	Diff		Env	Sup	Test	Std	Diff		Env	Sup	Test	Std	Diff	
IR 71692-45-2-2-1	IR 64	6	5	3.86	3.16	0.7	0.06	5	4	79.1	91.2	-12.1	0.11						
IR 71692-45-2-2-1	Azucena	5	4	4.2	3.7	0.5	0.11	5	5			-8.5	0.09						
IR 71692-45-2-2-1	Vandana	7	5			1.1	0.05	6	5			-10.3	0.05						
IR 71692-45-2-2-1	FR 13 A	4	4			0.1	0.21	4	3					5	4	R	MR	1	0.12
IR 71687-48-2-1-1	IR 64	6	5											6	5	MR	S	2	0.11
IR 71687-48-2-1-1	Azucena	4	3											4	4			1	0.09
IR 71687-48-2-1-1	Vandana	6	5											7	6			1	0.05
IR 71687-48-2-1-1	FR 13 A	3	2																
IR 71687-11-2-2-1	IR 64	6	5																
IR 71687-11-2-2-1	Azucena	7	5																
IR 71687-11-2-2-1	Vandana	4	3																
IR 71687-11-2-2-1	FR 13 A	3	3																
IR 71684-36-3-3-2	IR 64	6	5																
IR 71684-36-3-3-2	Azucena	4	3																
IR 71684-36-3-3-2	Vandana	5	5											5	4			1	0.12
IR 71684-36-3-3-2	FR 13 A	7	6											6	5			2	0.06

5. Stability Queries.

Find mean phenotypic data for a set of lines in each of several environments and obtain the mean performance across all lines grown in that environment.