New tools to improve efficiency of data analysis

Reinhard Simon, Raul Eyzaguirre, Luka Wanjohi,
Omar Benites, Luis Duque, Awais Khan

June, 2016

Nairobi, Kenya

SASHA Sweetpotato Action for Security and Health in Africa

BTI BOYCE THOMPSON INSTITUTE

CENTRO INTERNACIONAL DE LA PAPA CIP
A member of the CGIAR Consortium

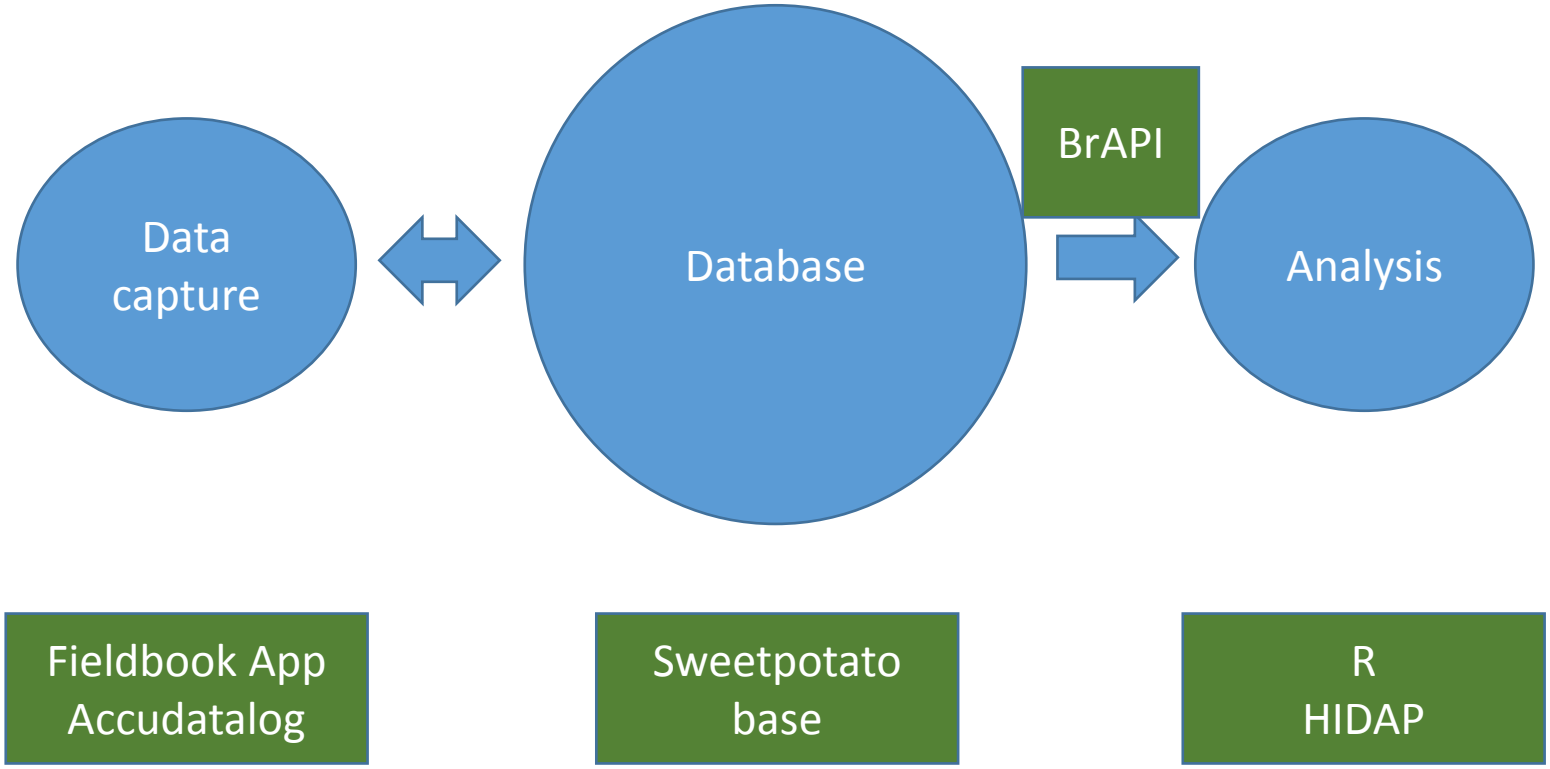
NC STATE UNIVERSITY

GT4SP

Motivation

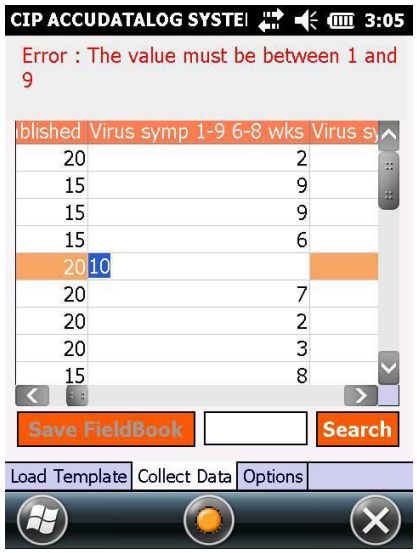
- Faster routine analysis of breeding trials
  - Building on current tools like CloneSelector
- Working with community tools:
  - Sweetpotatobase
  - KSU fieldbook app
  - Accu datalog
- Technical update to take better advantage of
  - Interactivity: Linked data & linked views
  - R reproducible reports for automating analyses
  - Ontologies to facilitate (statistical) handling of variables

The big picture …



AccuDataLog

- Mobile field data collection app.
- **Runs on Windows mobile for robust data collection in the field**
- **… and Android**
- **English, Swahili and Chinese**
- Developed by **CIP(SASHA)**
- Automatic Import of Fieldbooks into mobile device
- Field based data entry



Main Features

- Integrated barcode label technology (1D & 2D)
- Realtime data entry validation: numeric, date, string length, lower limits, upper limits, etc
- User defines lower and upper limits in excel
- Print on demand (POD) of barcode labels via mobile printing

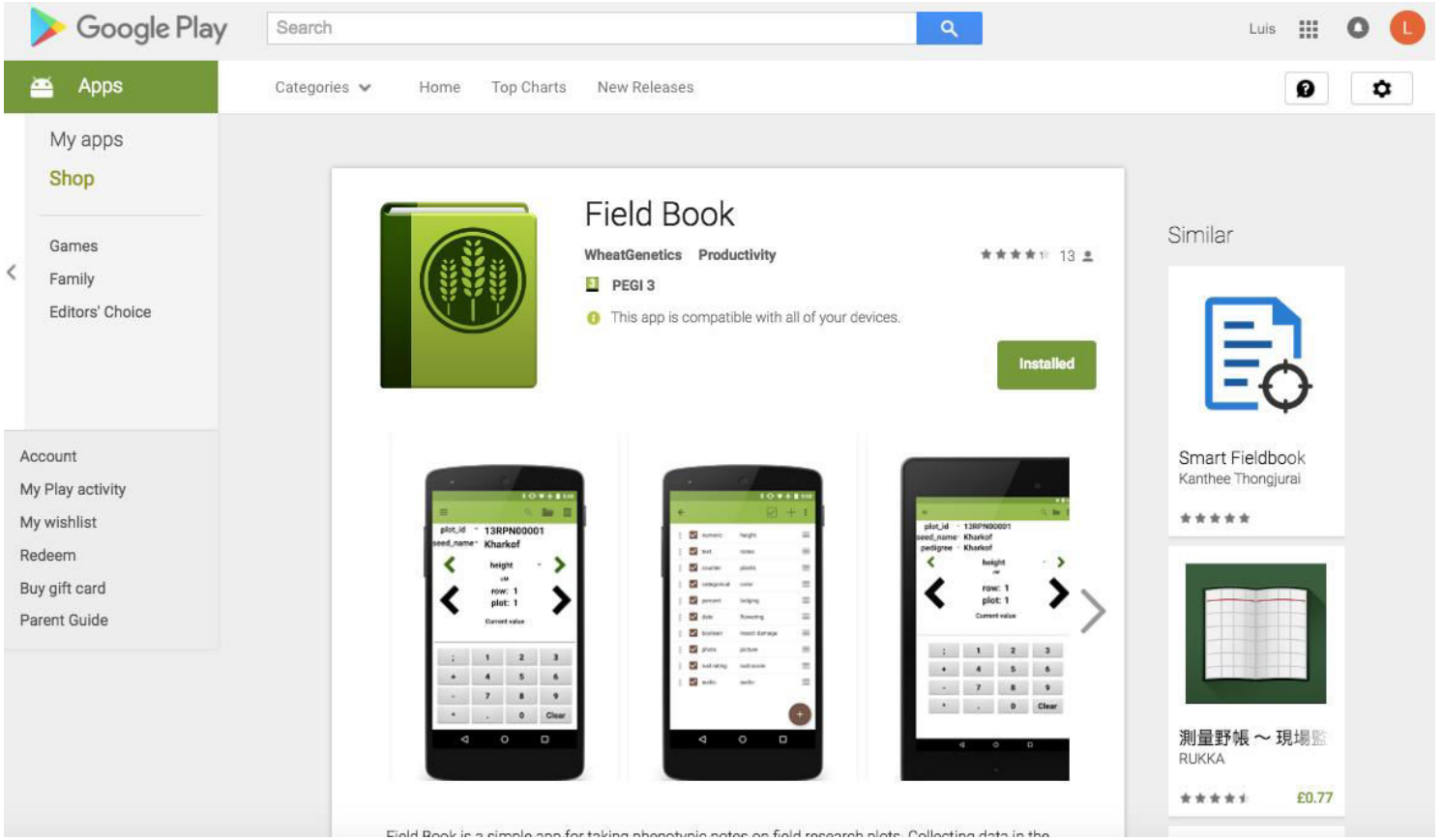
AccuDataLog Adoption

Ghana

Mozambique

Uganda

https://play.google.com/store/apps/details?id=com.fieldbook.tracker&hl=en




SAMSUNG
12:45 PM
Mon, 16 September
Swipe screen to unlock


SAMSUNG

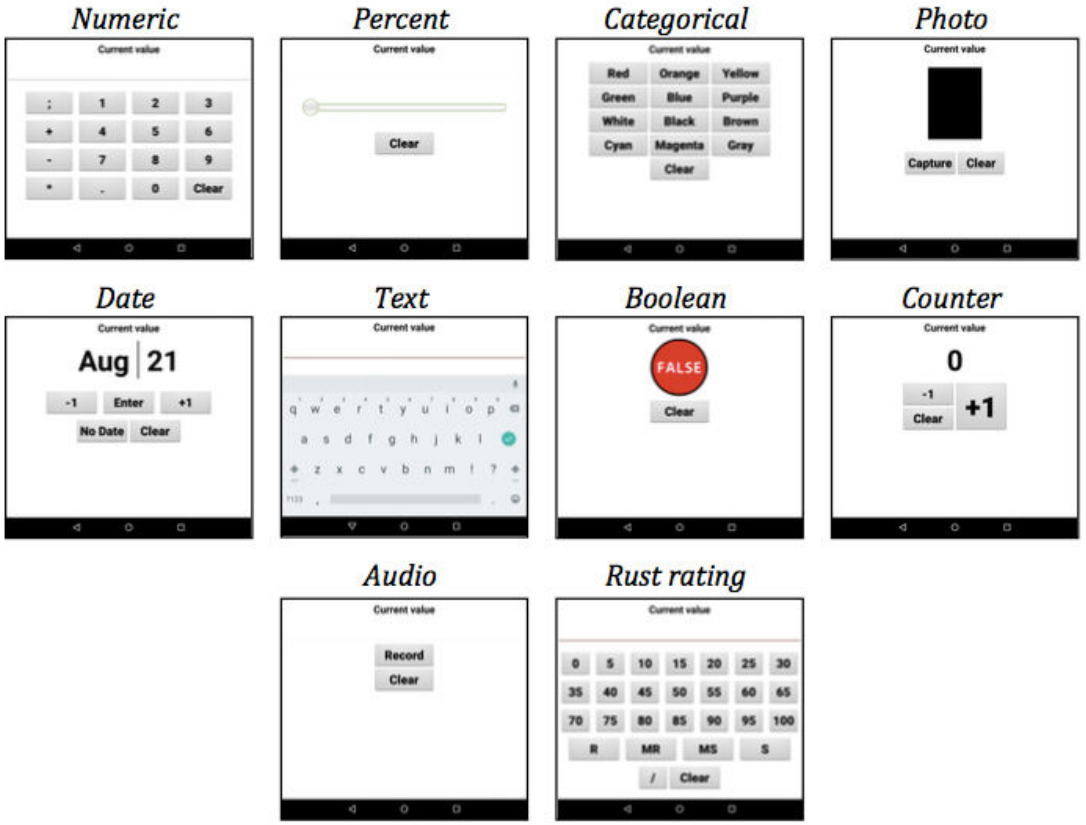

Otter
Otter


SanDisk
32GB
microSD
HC

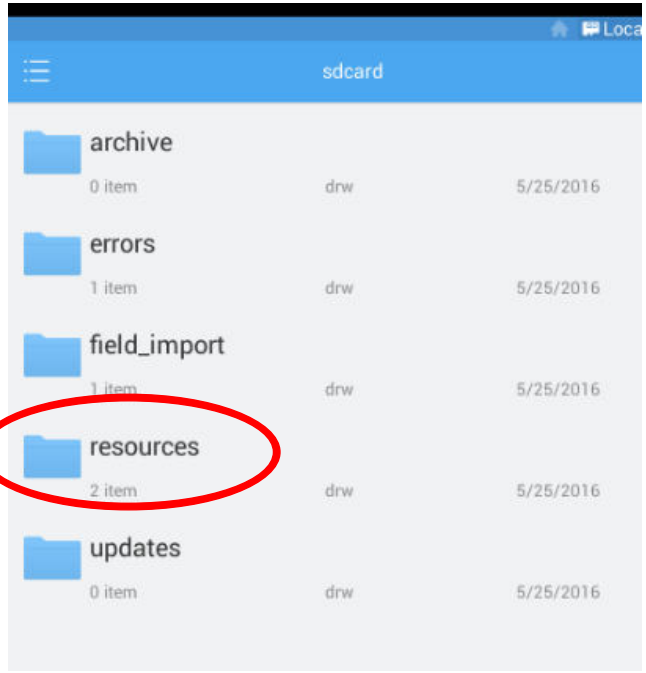

SAMSUNG
12:45 PM

Trait formats

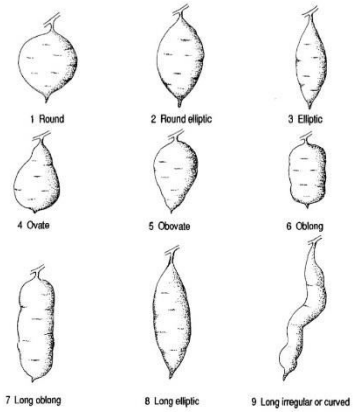
- Numeric
- Percent
- Categorical
- Date
- Text
- Boolean
- Counter
- Photo
- Audio

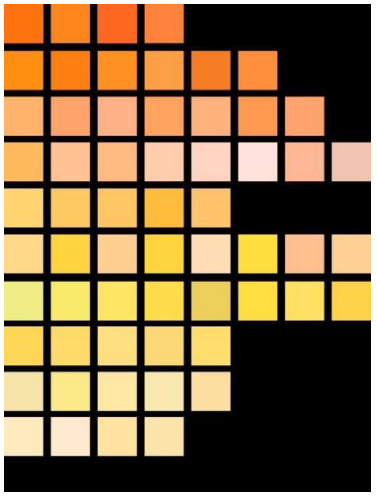


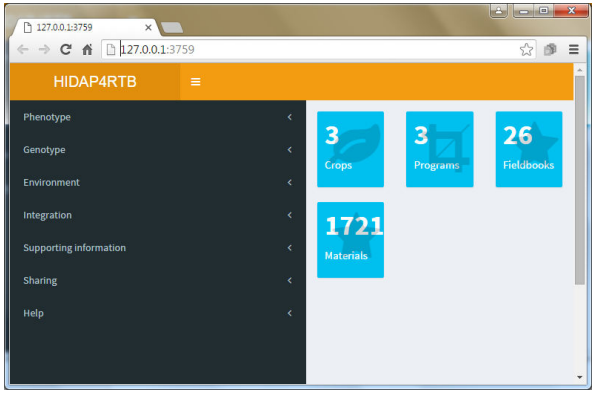
Resources Folder

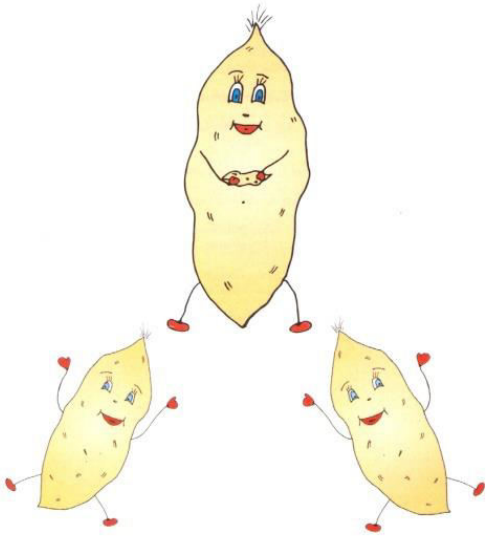


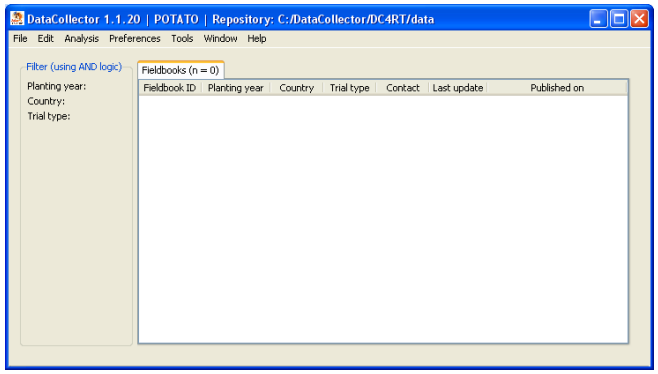
- Resources folder: user can upload cheat sheets, pictures, images, etc. to aid phenotyping

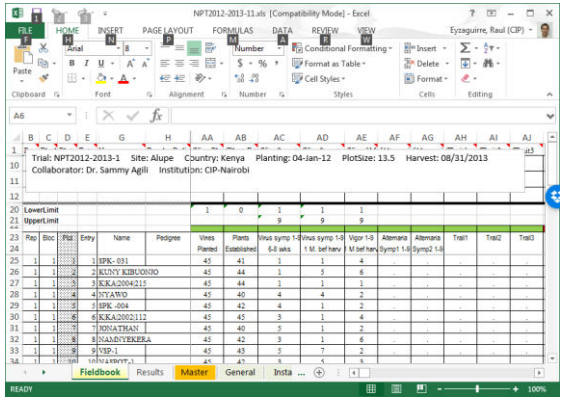












Data collector

Clone selector

HIDAP

Phenotype

Analysis

Fieldbook

SPYLAT2013_MZ-Gurue

Environment

Fieldbook

Show 10 entries

Search:

	PLOT	plotDbId	BLOCK	REP	germplasmDbId	germplasmName	Beta carotene content measuring mg per 100g	Fibers in cooked samples 1 estimating 1-9	Harvest index computing percent	Number of commercial storage roots counting number per plot	non-s num
73	1	38843	1	1	38837	MGSG 1065-4	11.03	1	40.1	19	
34	2	38845	1	1	38844	MGSG 1007-13	7.23	1	29	5	
55	3	38847	1	1	38846	MGSG 1012-9	7.23	1	62.7	14	
76	4	38849	1	1	38848	MGSG 1068-1	7.23	1	14.9	4	
10	5	38851	1	1	38850	MGSG 1001-7	3.03	3	58.7	10	
67	6	38852	1	1	38835	MGSG 1051-1	7.76	1	15	7	
19	7	38854	1	1	38853	MGSG 1004-27	7.23	1	66.9	12	
70	8	38856	1	1	38855	MGSG 1061-3	7.23	1	62.3	25	
37	9	38858	1	1	38857	MGSG 1007-9	6.12	1	8	4	
46	10	38860	1	1	38859	MGSG 1010-10	7.23	1	51.5	17	

Showing 1 to 10 of 81 entries

Previous 1 2 3 4 5 … 9 Next

HIDAP

≡

Phenotype

Analysis

Fieldbook

SPYLAT2013_MZ-Gurue

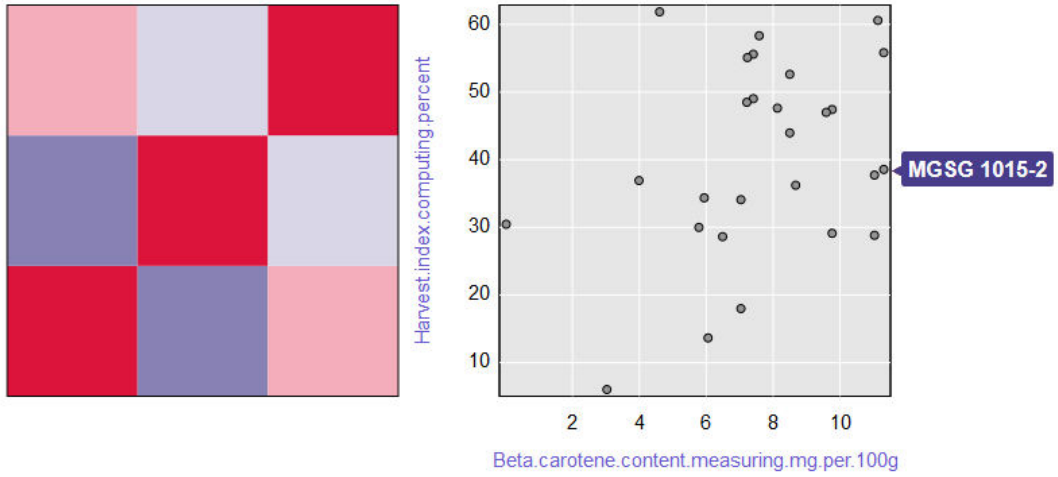
Environment

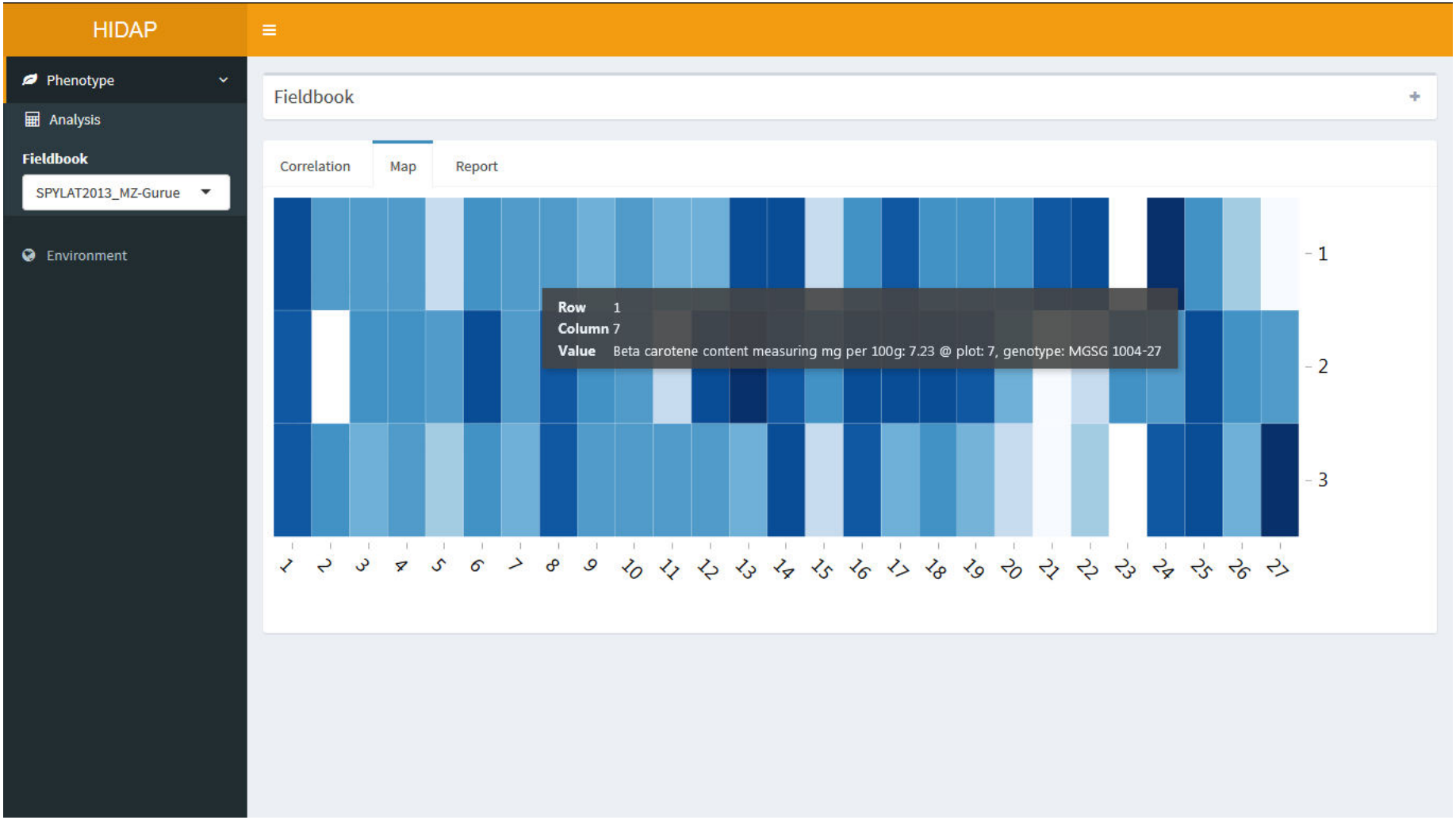
Fieldbook +

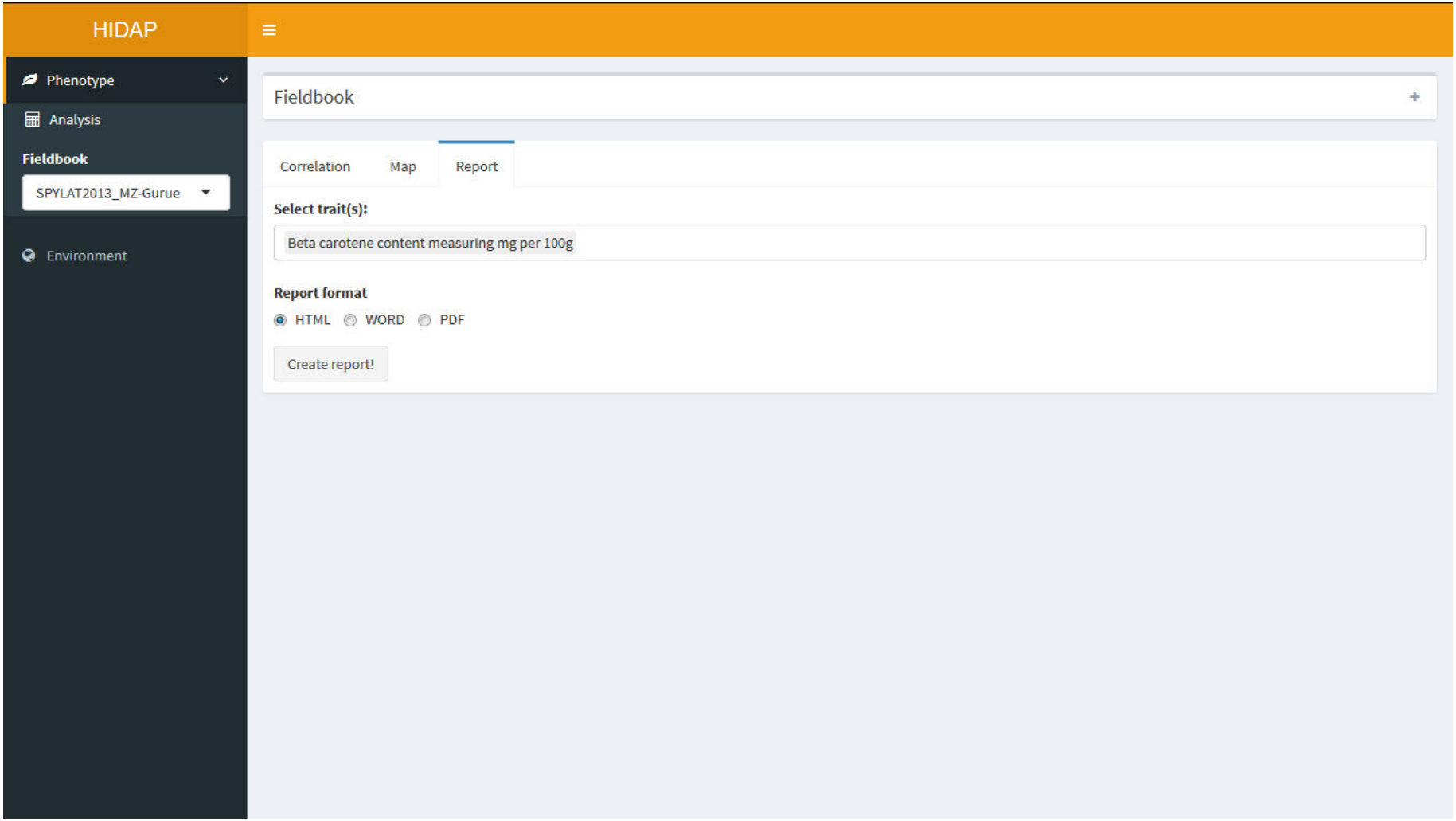
Correlation | Map | Report

Select two or more traits:

Beta carotene content measuring mg per 100g | Fibers in cooked samples 1 estimating 1-9 | Harvest index computing percent




HIDAP
Phenotype
Analysis
Fieldbook
SPYLAT2013_MZ-Gurue
Environment
Fieldbook
Correlation
Map
Report
Row 1
Column 7
Value Beta carotene content measuring mg per 100g: 7.23 @ plot: 7, genotype: MGSG 1004-27
1
2
3
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27


HIDAP
Phenotype
Analysis
Fieldbook
SPYLAT2013_MZ-Gurue
Environment
Fieldbook
Correlation
Map
Report
Select trait(s):
Beta carotene content measuring mg per 100g
Report format
HTML
WORD
PDF
Create report!

ANOVA for a RCBD trial: SPYLAT2013_MZ-Gurue

HIDAP

June 07, 2016, 10:12h


```
# This is an automatedly created report.

# See more details in section on materials.
```

Abstract

This trial has the identifier SPYLAT2013_MZ-Gurue. It was conducted under the supervision of x y as a Advanced Trial as part of a Yield Breeding Program in Gurue, Mozambique, Z in 2016. A total of 27 clones (including reference clones) were evaluated for 1 traits.

Materials and Methods

Model specification and data description

There is data from 27 treatments, evaluated using a randomize complete block design with 1, 2, 3 blocks. The statistical model is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

where

- y_{ij} is the observed response with treatment i and block j.
- μ is the mean response over all treatments and blocks.
- τ_i is the effect for treatment i.

Trait analyses

The following traits were not analyzed since they had too many missing values (>= 10%): . For the remaining traits missing values were imputed using all available information.

Valid traits: **Beta carotene content measuring mg per 100g**.

Analysis of **Beta carotene content measuring mg per 100g**

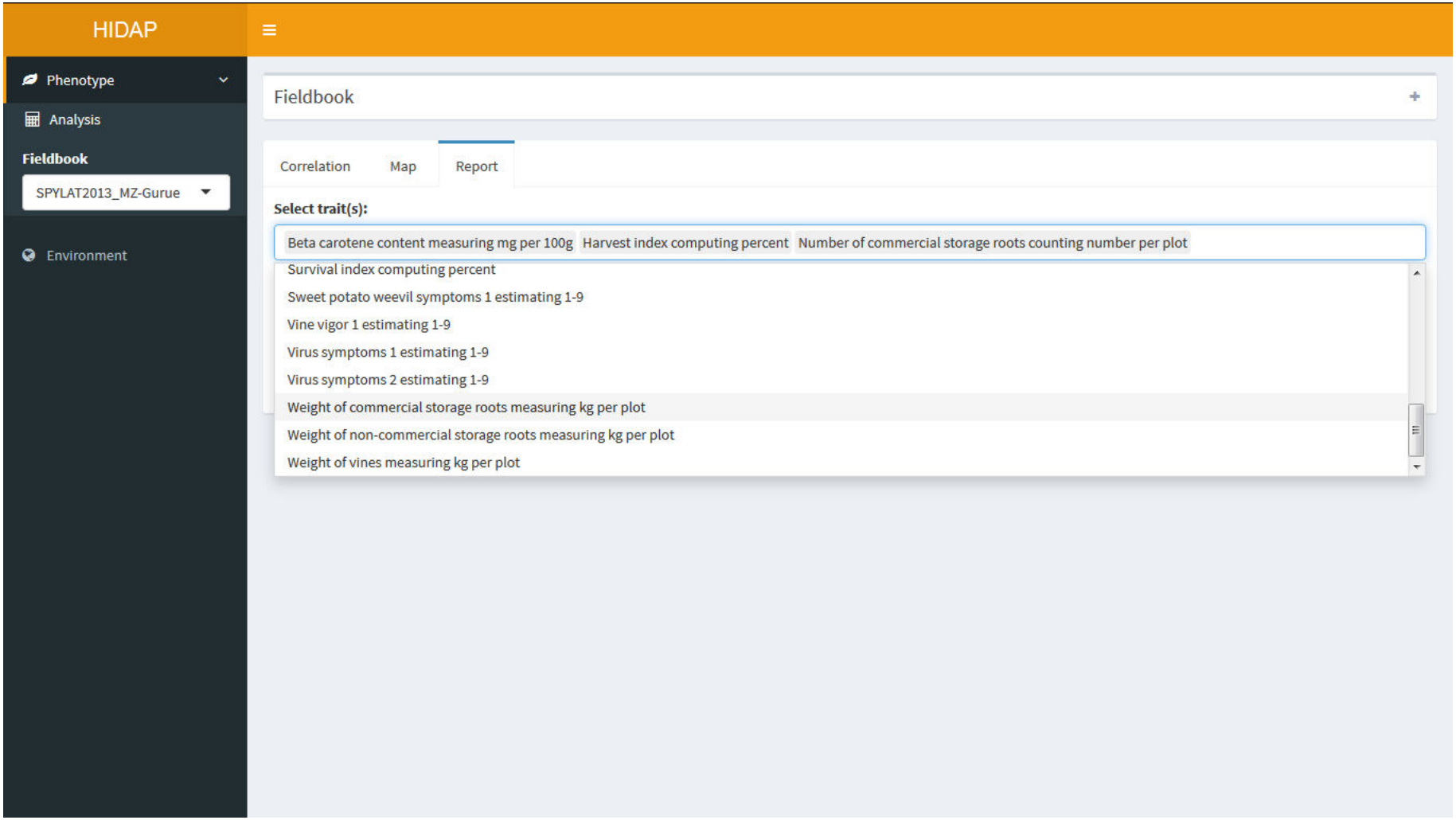
You have fitted a linear model for a RCBD. The ANOVA table for your model is:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
germplasmName	26	562.042	21.617	16.5482	5.16033e-17
REP	2	16.3471	8.17357	6.25703	0.00367343
Residuals	52	67.9276	1.3063	NA	NA

The p-value for treatments is 0.00000000000000000516033 which is significant at the 5% level.

The means of your treatments are:

germplasmName	Beta carotene content measuring mg per 100g
Chingova	0.02
Jonathan	3.99
MGSG 1001-36	11.1
MGSG 1001-7	4.61
MGSG 1002-49	8.5
MGSG 1003-27	7.73
MGSG 1004-2	9.76
MGSG 1004-27	7.41
MGSG 1005-17	5.94


HIDAP
Phenotype
Analysis
Fieldbook
SPYLAT2013_MZ-Gurue
Environment
Fieldbook
Correlation
Map
Report
Select trait(s):
Beta carotene content measuring mg per 100g
Harvest index computing percent
Number of commercial storage roots counting number per plot
Survival index computing percent
Sweet potato weevil symptoms 1 estimating 1-9
Vine vigor 1 estimating 1-9
Virus symptoms 1 estimating 1-9
Virus symptoms 2 estimating 1-9
Weight of commercial storage roots measuring kg per plot
Weight of non-commercial storage roots measuring kg per plot
Weight of vines measuring kg per plot

ANOVA for a RCBD trial: SPYLAT2013_MZ-Gurue

HIDAP

June 07, 2016, 10:13h

Contents


```
# This is an automatedly created report.

# See more details in section on materials
```

Analysis of Harvest index computing percent

You have fitted a linear model for a RCBD. The ANOVA table for your model is:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
germplasmName	26	16327	627.961	5.16952	2.46623e-07
REP	2	101.453	50.7264	0.417592	0.66082
Residuals	52	6316.63	121.474	NA	NA

The p-value for treatments is 0.000000246623 which is significant at the 5% level.

The means of your treatments are:

germplasmName	Harvest index computing percent
Chingova	30.5
Jonathan	36.9
MGSG 1001-36	60.6
MGSG 1001-7	61.9
MGSG 1002-49	52.6
MGSG 1003-27	30
MGSG 1004-2	47.4
MGSG 1004-27	55.6
MGSG 1005-17	34.4
MGSG 1006-7	47
MGSG 1006-9	34.1
MGSG 1007-13	47.6
MGSG 1007-9	13.7
MGSG 1008-8	28.6
MGSG 1009-3	58.3
MGSG 1010-10	49
MGSG 1010-4	55.8
MGSG 1011-5	48.5
MGSG 1012-9	44
MGSG 1015-17	28.8
MGSG 1015-2	38.6

Model specification and data description

There is data from 27 treatments, evaluated using a randomize complete block design with 1, 2, 3 blocks. The statistical model is

$$y_{ij} = \mu + \tau_i + \beta_j + \epsilon_{ij}$$

where

- y_{ij} is the observed response with treatment i and block j.
- μ is the mean response over all treatments and blocks.
- τ_i is the effect for treatment i.
- β_j is the effect for block j.
- ϵ_{ij} is the error term.

In this model we assume that the errors are independent and have a normal distribution with common variance, that is, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$.

The following traits are analyzed: **Beta carotene content measuring mg per 100g, Harvest index computing percent, Weight of commercial storage roots measuring kg per plot.**

The following germplasm was analyzed: MGSG 1065-4, MGSG 1007-13, MGSG 1012-9, MGSG 1068-1, MGSG 1001-7, MGSG 1051-1, MGSG 1004-27, MGSG 1061-3, MGSG 1007-9,

Analysis of Beta carotene content measuring mg per 100g

You have fitted a linear model for a RCBD. The ANOVA table for your model is:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
germplasmName	26	562.082	21.6186	16.3243	6.95787e-17
REP	2	17.1183	8.55914	6.46305	0.00311304
Residuals	52	68.8645	1.32432	NA	NA

The p-value for treatments is 0.000000000000000695787 which is significant at the 5% level.

The means of your treatments are:

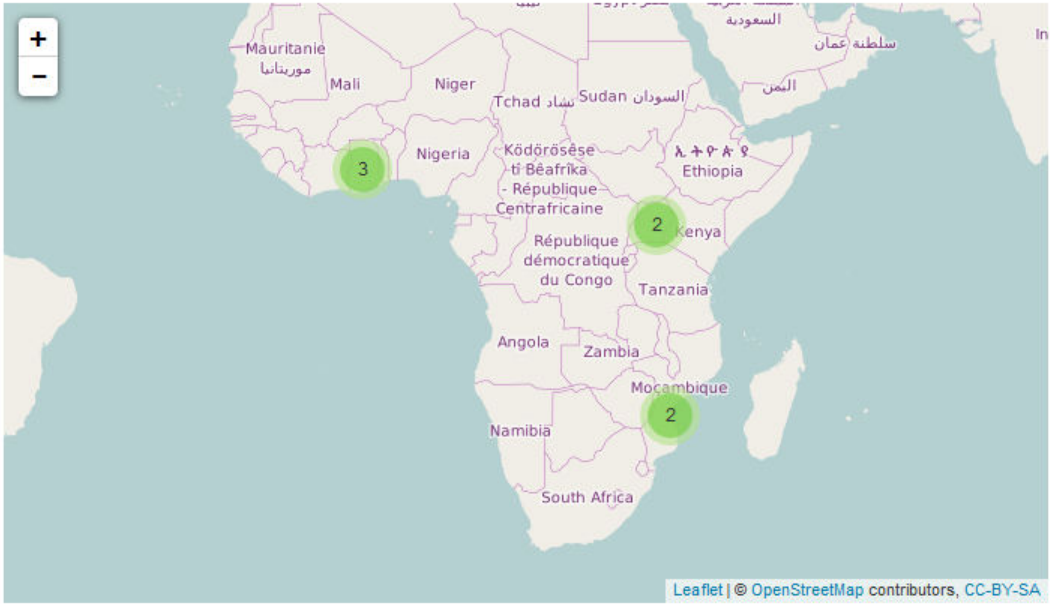
germplasmName	Beta carotene content measuring mg per 100g
Chingova	0.02
Jonathan	3.99
MGSG 1001-36	11.1
MGSG 1001-7	4.61
MGSG 1002-49	8.5
MGSG 1003-27	7.81
MGSG 1004-2	9.76
MGSG 1004-27	7.41
MGSG 1005-17	5.94
MGSG 1006-7	9.59
MGSG 1006-9	7.04
MGSG 1007-13	8.13
MGSG 1007-9	6.05
MGSG 1008-8	6.49
MGSG 1009-3	7.58

HIDAP

Phenotype

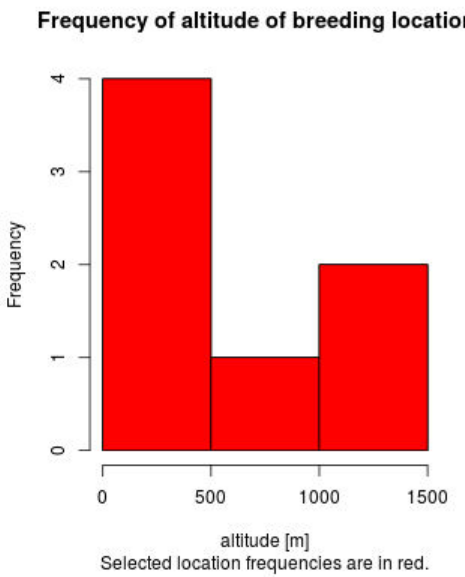
Environment

Map | Report



Histogram | Info | Fieldtrials

Genotypes | Site



Location table

Show 10 entries | Search:

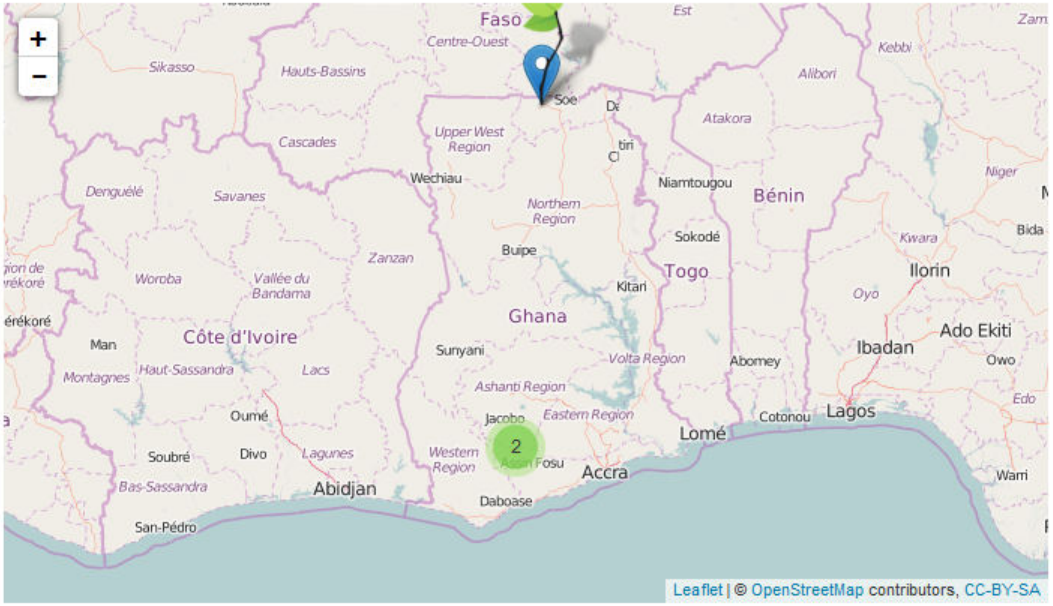
	locationDbId	altitude	countryCode	countryName	name	longitude
1	10				Cornell Biotech	

HIDAP

Phenotype

Environment

Map | Report



Location table

Show 10 entries Search:

	locationDbId	altitude	countryCode	countryName	name	longitude
1	10				Cornell Biotech	
2	12				Clinton, NC	

Histogram | Info | Fieldtrials

Genotypes Site

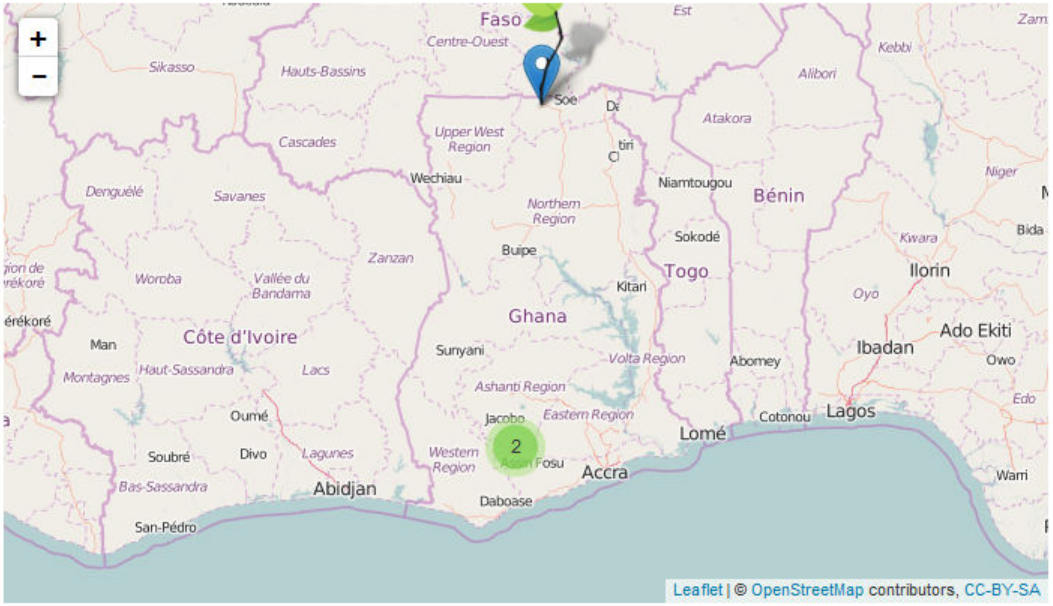
	Attribute	Value
1	locationDbId	18
5	name	Tono, Ghana
9	Uniquename	
4	countryName	
3	countryCode	
2	altitude	183
6	longitude	-1.1466
7	latitude	10.87
8	geodetic.datum	
10	Agricultural.Ecological.Zone	
12	adm1	
13	adm2	
14	adm3	
11	Continent	

HIDAP

Phenotype

Environment

Map | Report



Histogram | Info | Fieldtrials

Genotypes

Site

SPYLPT2013_GH-Tono

Location table

Show 10 entries

Search:

	locationDbId	altitude	countryCode	countryName	name	longitude
1	10				Cornell Biotech	
2	12				Clinton, NC	

Trial detail for SPYLPT2013_GH-Tono

Trial details

Breeding Program	Ghana (Ghana)	
Trial Name	SPYLPT2013_GH-Tono	[change]
Trial Type	PYT	[change]
Year	2013	[change]
Trial Location	Tono, Ghana	[change]
Planting Date	[No Planting Date]	[change]
Harvest Date	[No Harvest Date]	[change]
Description	163 sweetpotato clones were assayed in this Preliminary Trial of a yield breeding program for 25 traits at Tono	[edit]

Folder [New Folder] | [Change]

Ghana

HIDAP Trial Analysis

Phenotype Plot Map | Phenotype Correlation | Analysis Report

Physical Trial Layout

[Upload trial coordinates]

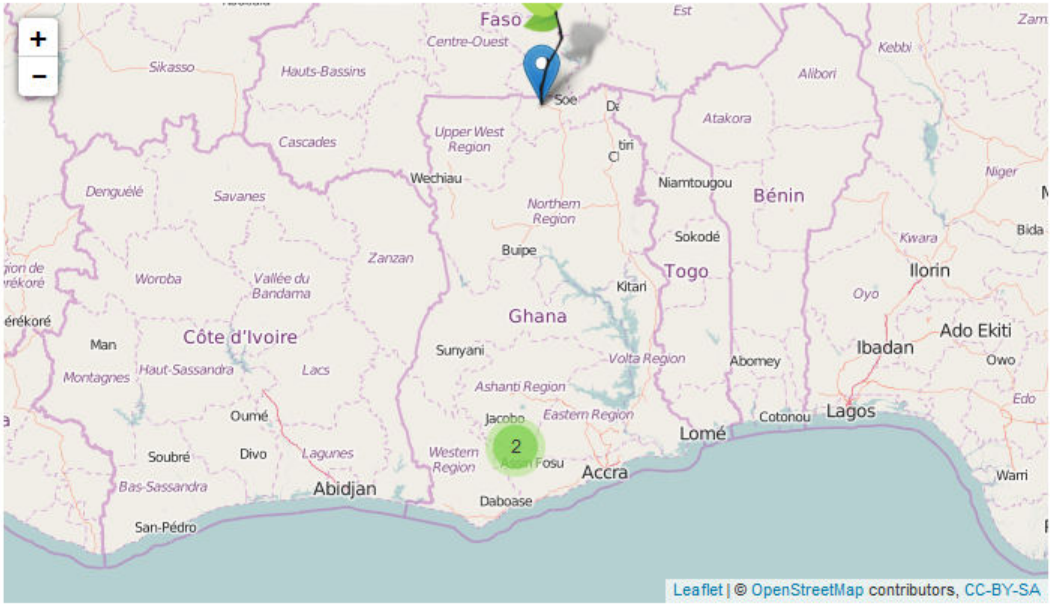
HIDAP

≡

Phenotype

Environment

Map | Report



Location table

Show 10 entries Search:

	locationDbId	altitude	countryCode	countryName	name	longitude
1	10				Cornell Biotech	
2	12				Clinton, NC	

Histogram | Info | Fieldtrials

Genotypes

Site

Top genotypes for trait (Harvest index) from most recent (2013) fieldbook: SPYLPT2013_GH-Tono for location: :
PGB12162-4 (64.3), PG12170-4 (63.2), PG12151-26 (57.3), Apomuden (56.8)

Accession: Apomuden

Stock details New QTL population | Back to stock search

[New] [Edit] [Delete]

Organism	**Ipomoea batatas**
Stock type	**accession**
Stock name	**Apomuden**
Uniquename	**Apomuden**
Description	

SPB stock 40582 (Apomuden)

SPB40582

Stock editors: Reinhard Simon

Synonyms [Add]

Additional information [Add]

Associated loci (0)

Experimental data

Pedigree [Add parent] [Remove parent]

Descendants None

Pedigree string

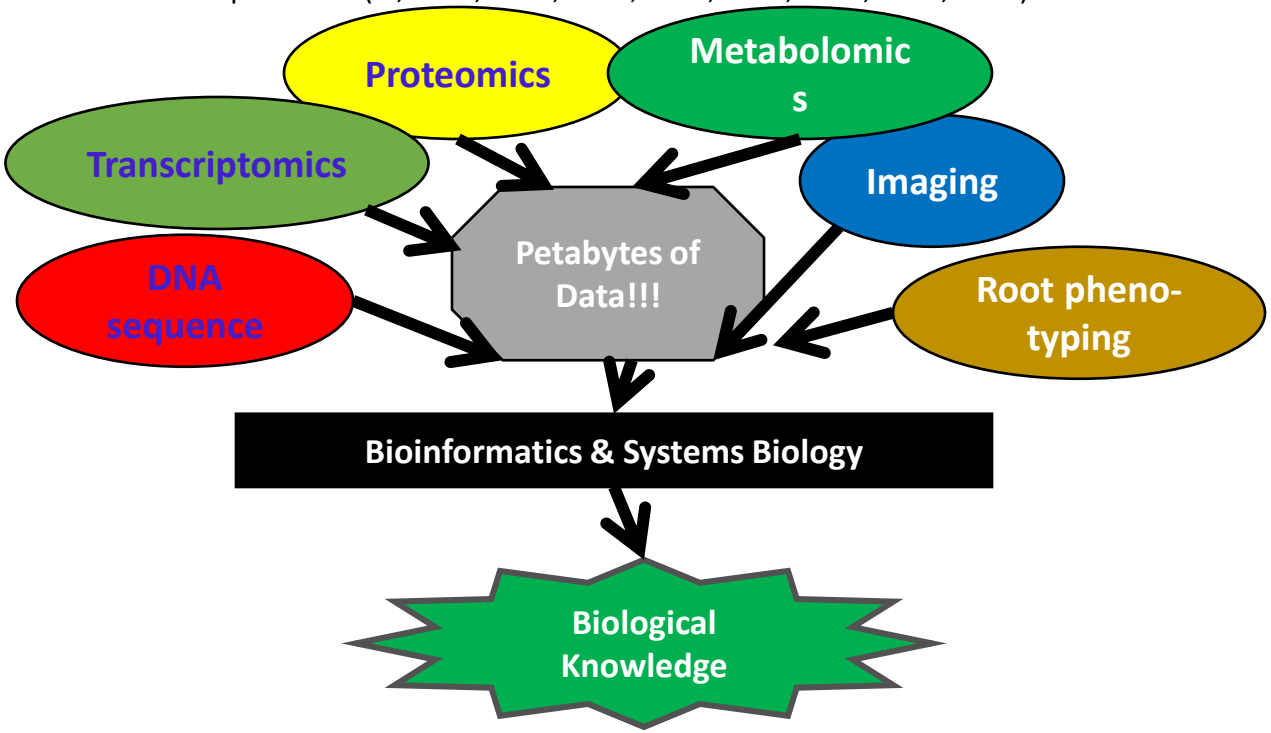
Parents

In the pipeline …

- Version 1.0 windows installer
- MET view
- Genotype data

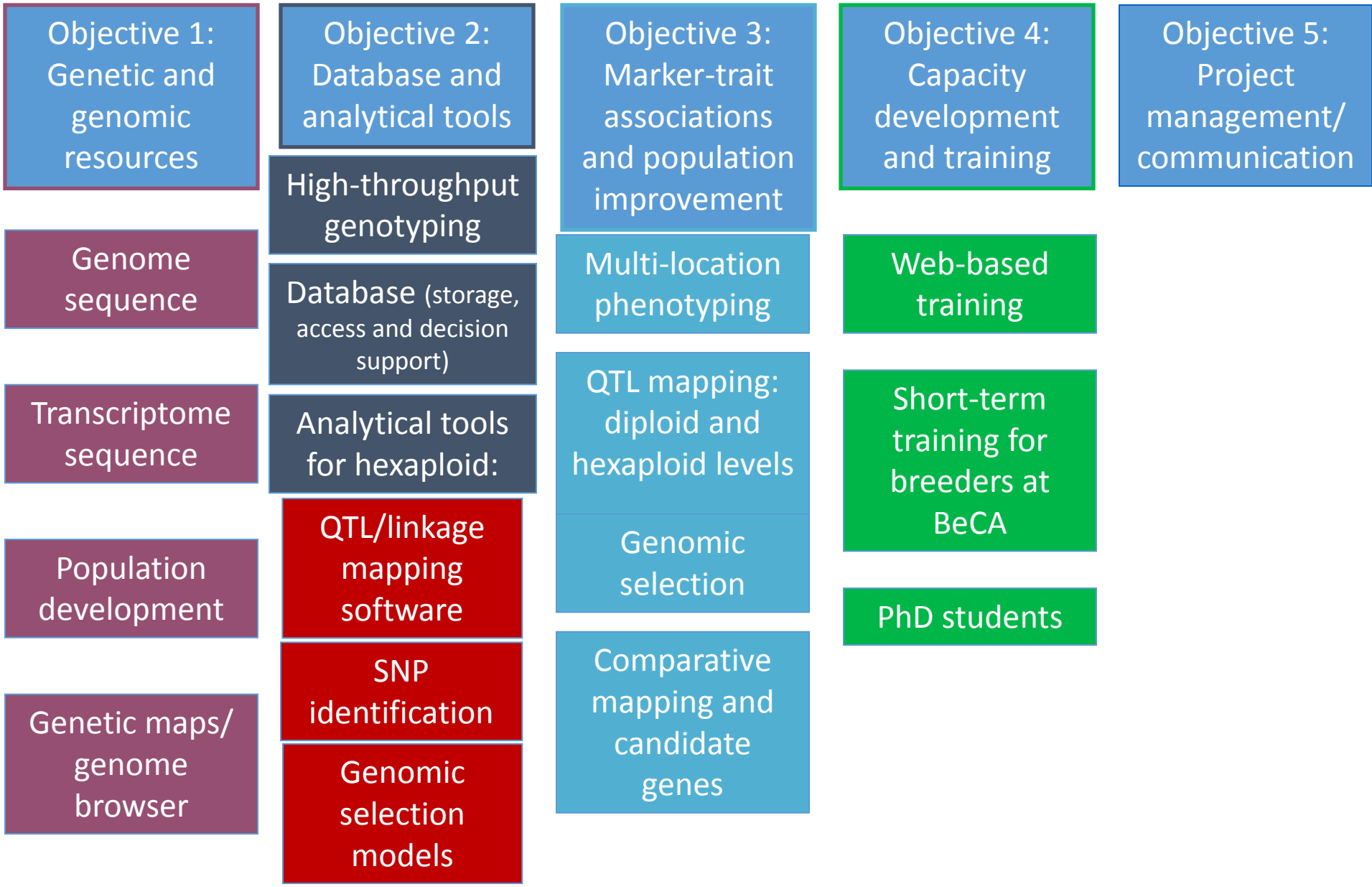
Integrated Analysis

The problem: improved genome sequencing and phenotyping have generated an enormous amount of data. In the near future, yottabytes of data can be expected (1,000,000,000,000,000,000,000,000)!!!



Outlook: new tools are needed for processing all this data, and integrating it for actual biological insights.

GT4SP Objectives



HIDAP: Next steps

- Stabilize the version with functionalities so far

- **Implement additional functionalities and tools developed by GT4SP colleagues at NCSU, UQ, MSU and BTI for integrated analysis of next generation multi-dimensional data sets** (QTL mapping, association mapping, selection schemes (MAS and GS))