Searching for traits of resistance to crop diseases and pests in plant genetic resources

Presented by Dag Endresen, GBIF-Norway, NHM-UiO

for the Nordic OIKOS symposium in Stockholm, 3rd February 2014

TOPICS:

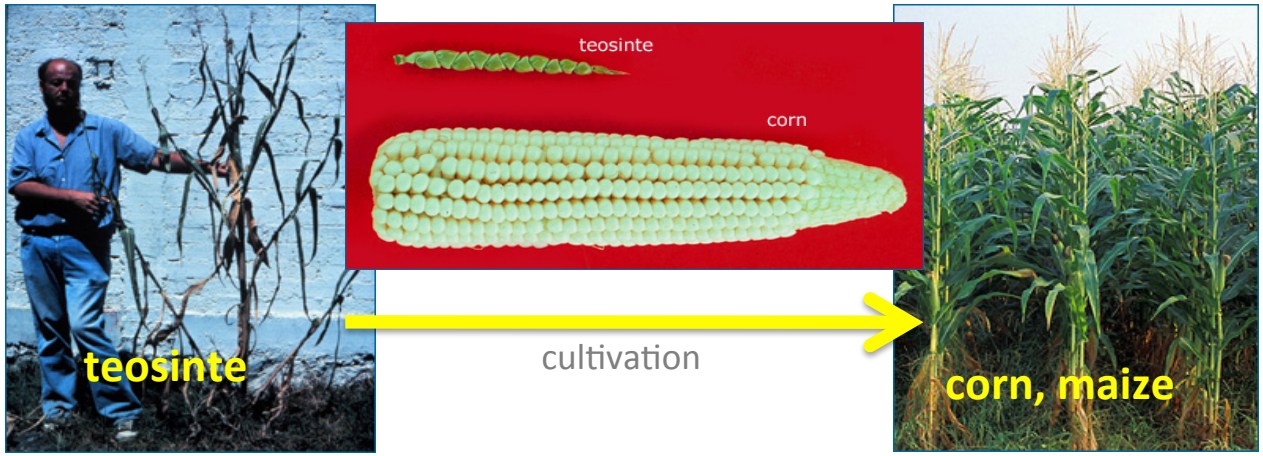
- Agro-ecology and trait mining
  - Predictive link between climate data and trait data
  - Some examples of studies:
    - Morphological traits in Nordic barley
    - Net blotch on barley landraces
    - Stem rust on wheat landraces
    - Stem rust, Ug99 on bread wheat

- Agricultural biodiversity informatics and GBIF

Wheat at Alnarp, June 2010

Domestication and cultivated plants:

Utilizing genetic potential from the wild

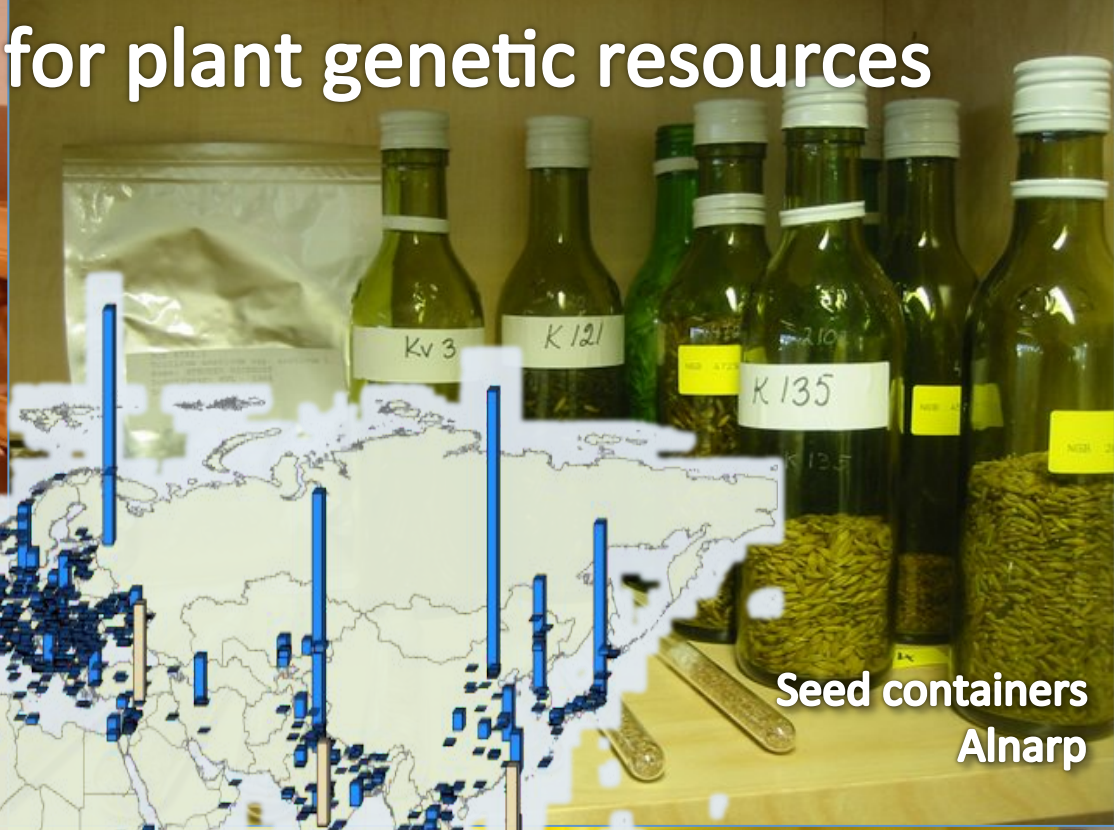


Centers of origin for selected crops (by USDA)

This map was developed by the United States Department of Agriculture (USDA).

Ex situ genebank collections for plant genetic resources

Seed drying room
Alnarp

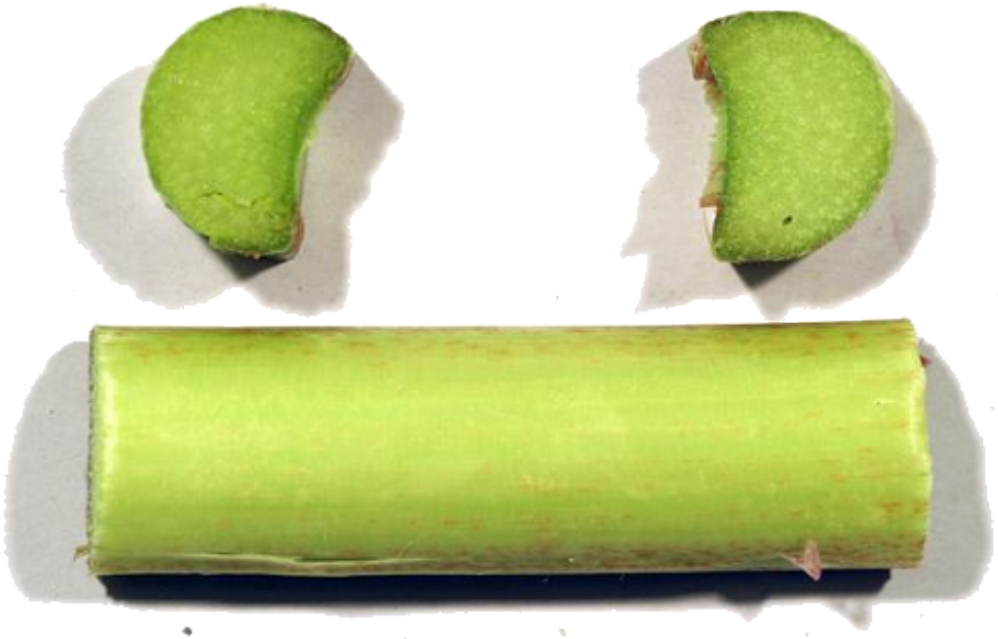


Seed containers
Alnarp

Svalbard Global
Seed Vault

Musa *in vitro*
Leuven

Conserving genetic diversity of agricultural traits



Rhubarb, *Rheum x hybridum* Murray, DKRHE43, by Gitte K. Bjørn

Field trials in Horsham Australia for cold tolerance in chickpea (*Cicer arietinum* L.). CC-by D. Endresen

Powdery Mildew,
Blumeria graminis

Leaf spots
Ascochyta sp.

Yellow rust
Puccinia strilformis

Black stem rust
Puccinia graminis

Apple, *Malus domestica* Borkh.
Cultivar "*Nanna*" by Stein H Hjeltnes

Field of sugar beet (*Beta vulgaris* L.) at Alnarp (June 2005) by Dag Endresen, CC-by.

A needle in a hay stack

- Scientists and plant breeders want a few hundred germplasm accessions to evaluate for a particular trait.

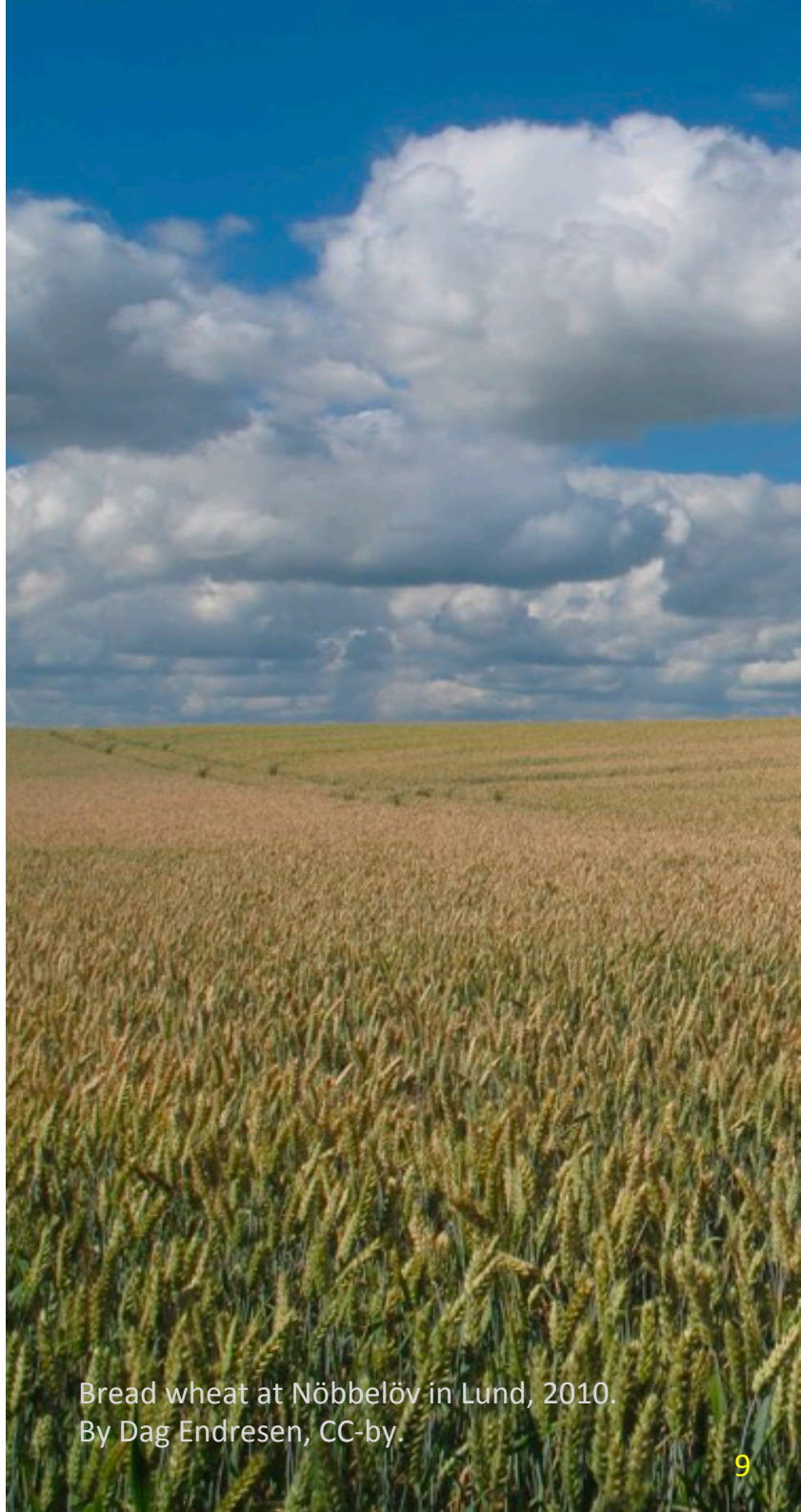
- How does the scientist select a small subset likely to have the useful trait?

Target:

– Identification of plant germplasm with a higher likelihood of having desired genetic diversity for a target trait property.

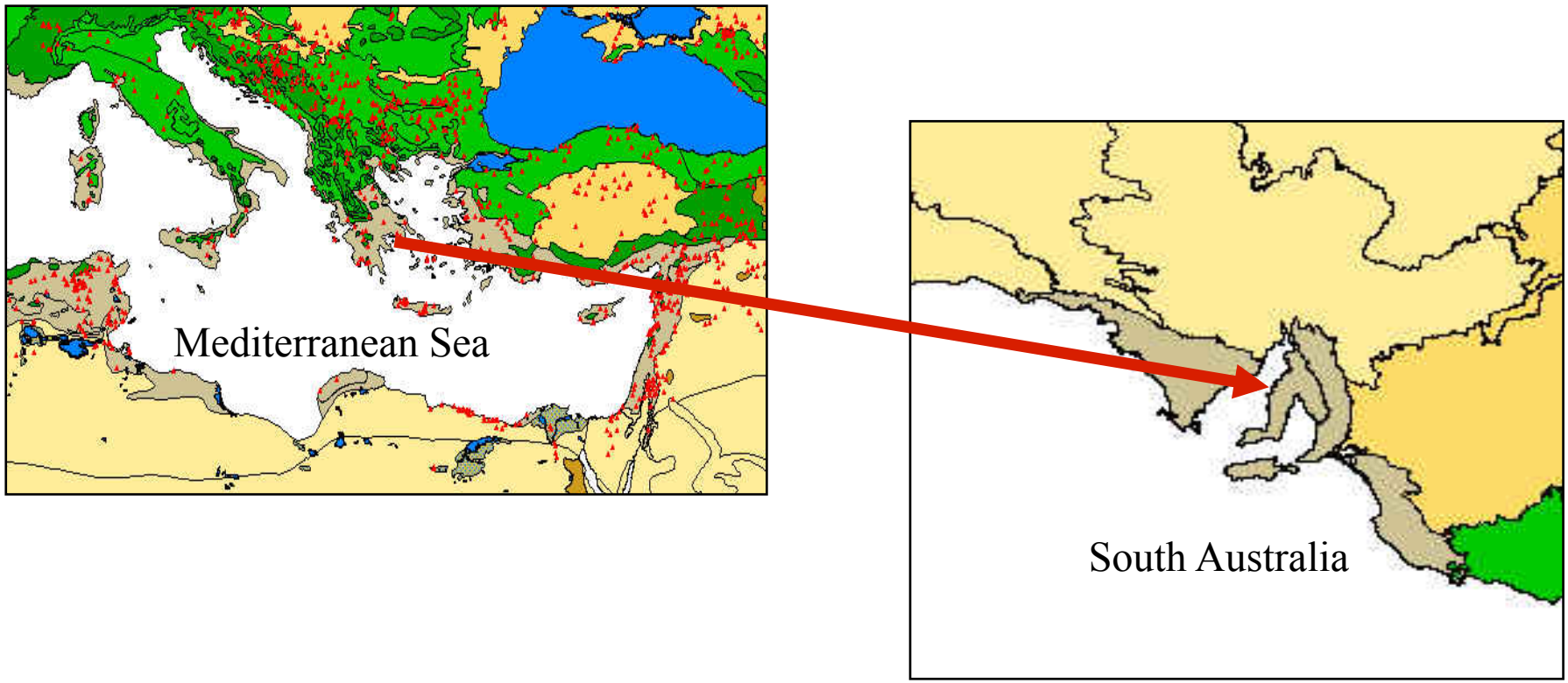
Suggested solution:

– Using environment data layers for predictive characterization of crop traits *a priori* BEFORE field trials.



Bread wheat at Nöbbelöv in Lund, 2010.
By Dag Endresen, CC-by.

Focused Identification of Germplasm Strategy



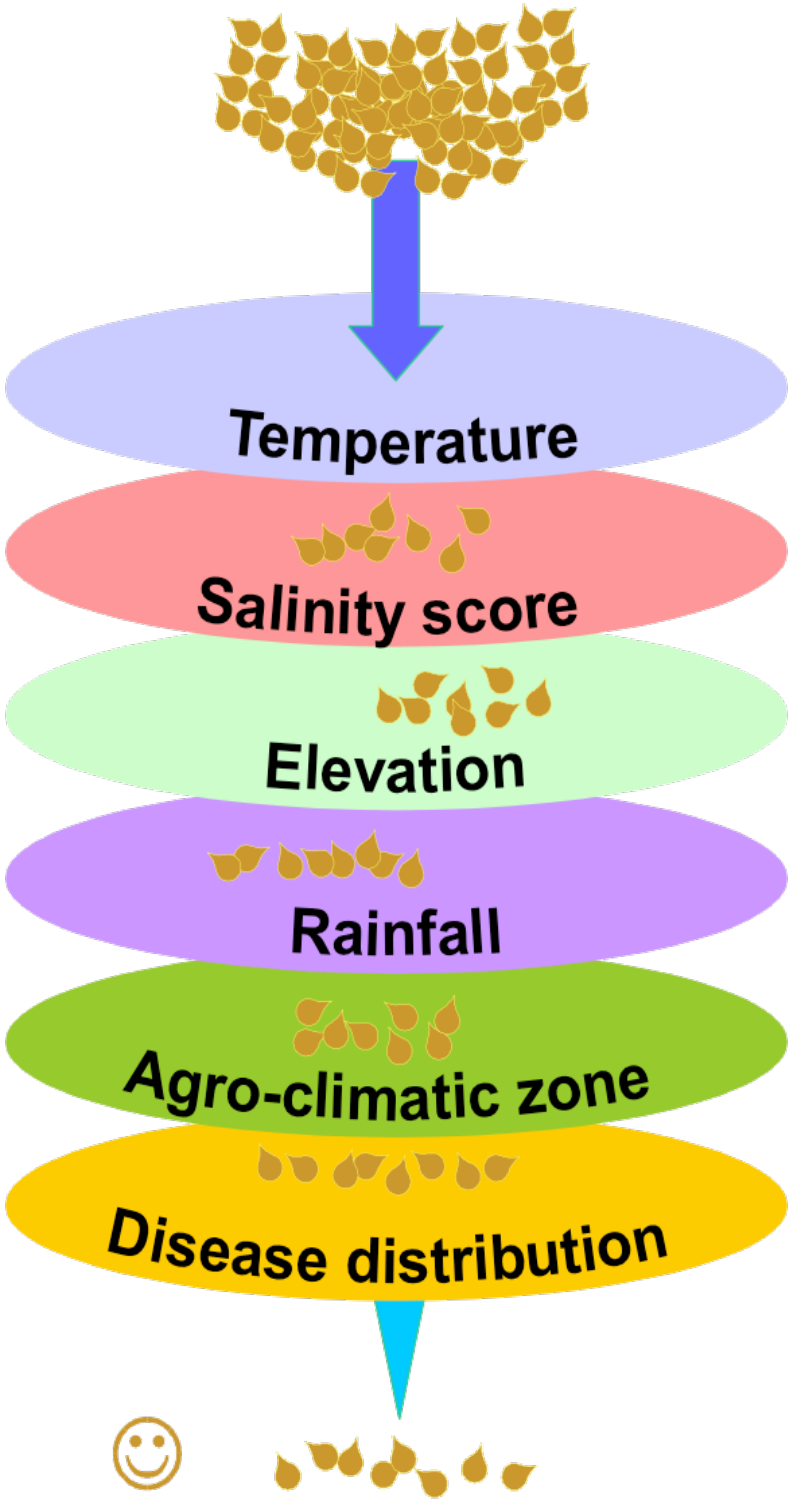
Origin of Concept:

Boron toxicity in marine sediments for wheat and barley genetic resources - example of late 1980s.

Michael Mackay (1986, 1990, 1995)

Focused Identification of Germplasm strategy

- Identify new and useful genetic diversity for crop improvement.
  - *Heuristic experience and expert knowledge for finding upper and lower limits for environmental variables (1).*
  - *Predictive eco-geographic data modelling analysis (2).*



http://www.figstraitmine.org/

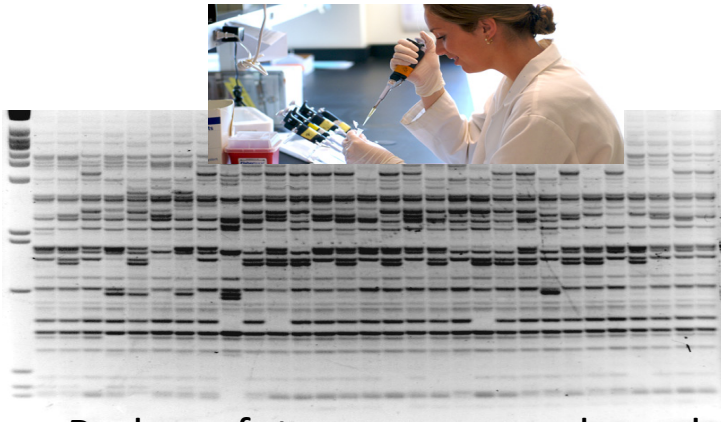
Climate effect during the breeding process

Wild relatives are shaped by the environment.

Primitive cultivated crops are shaped by local climate and humans.

Traditional cultivated crops (landraces) are shaped by climate and humans.

Modern cultivated crops are mostly shaped by humans (plant breeders).



Perhaps future crops are shaped in the molecular laboratory…?

Predictive link between eco-geography and traits

During traditional cultivation the farmer will select for and introduce germplasm for improved suitability of the landrace to the local conditions.

Limitations of FIGS

- Landraces and wild relatives
  - The link between climate data and the trait data is required for trait mining with FIGS. Modern cultivars are not expected to show this predictive link (complex pedigree).

- Georeferenced accessions
  - Trait mining with FIGS is based on multivariate models using climate data from the source location of the germplasm. To extract climate data the accessions need to be accurately georeferenced.

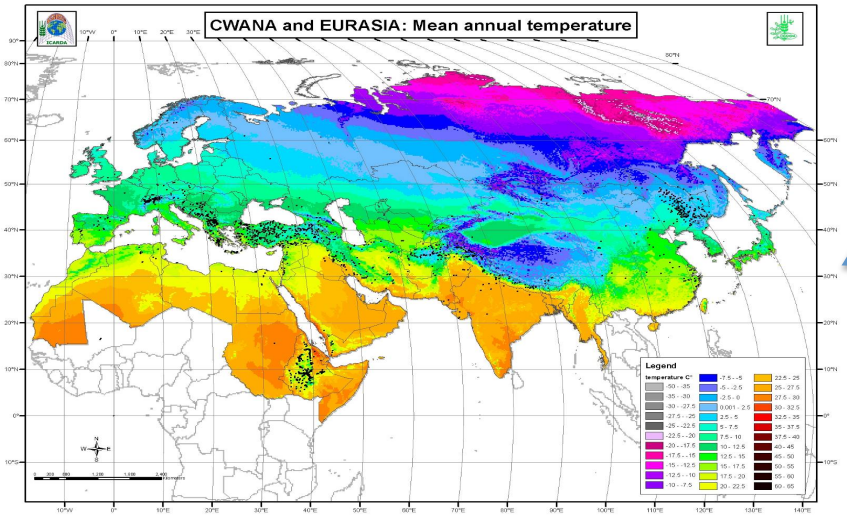
Wheat in the Hulah valley (Israel), 2007 by Aviad Bublil

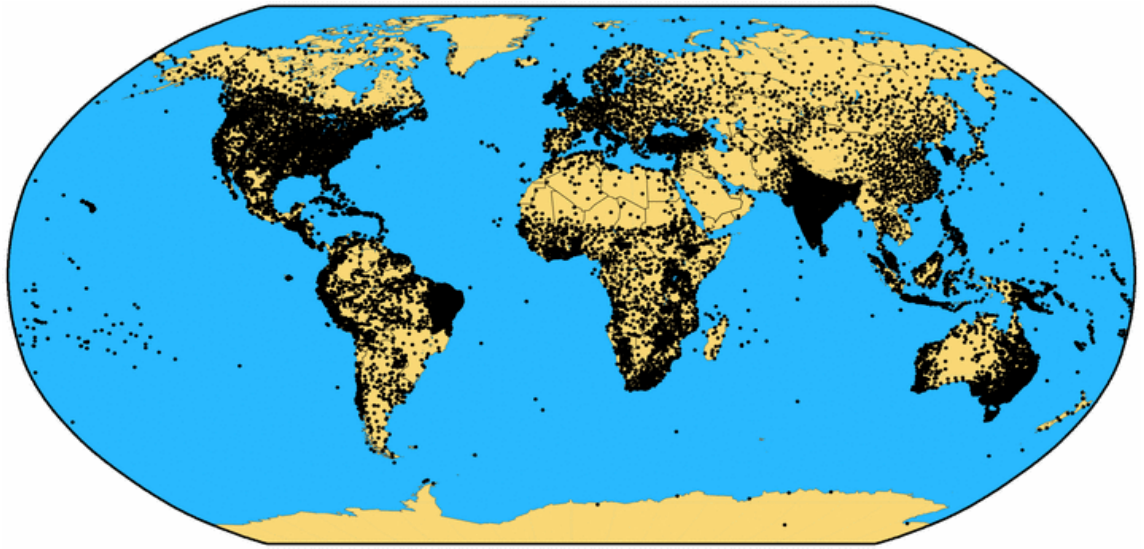
Climate data – WorldClim

The climate data can be extracted from the WorldClim dataset.
http://www.worldclim.org/
(Hijmans *et al*., 2005)

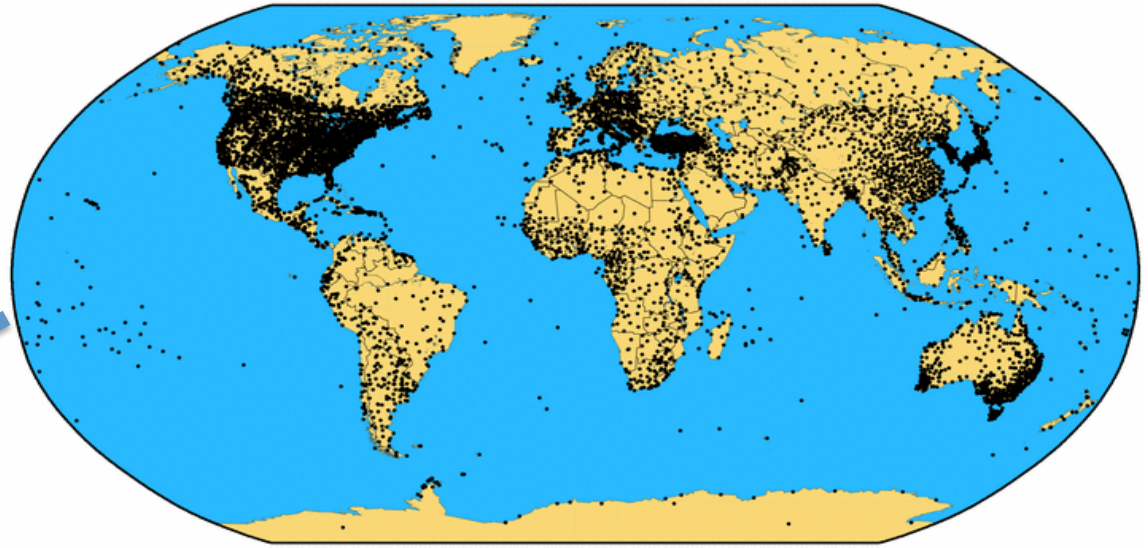
Data from weather stations worldwide are combined to a continuous surface layer.

Climate data for each landrace is extracted from this surface layer.





Precipitation: 20 590 stations



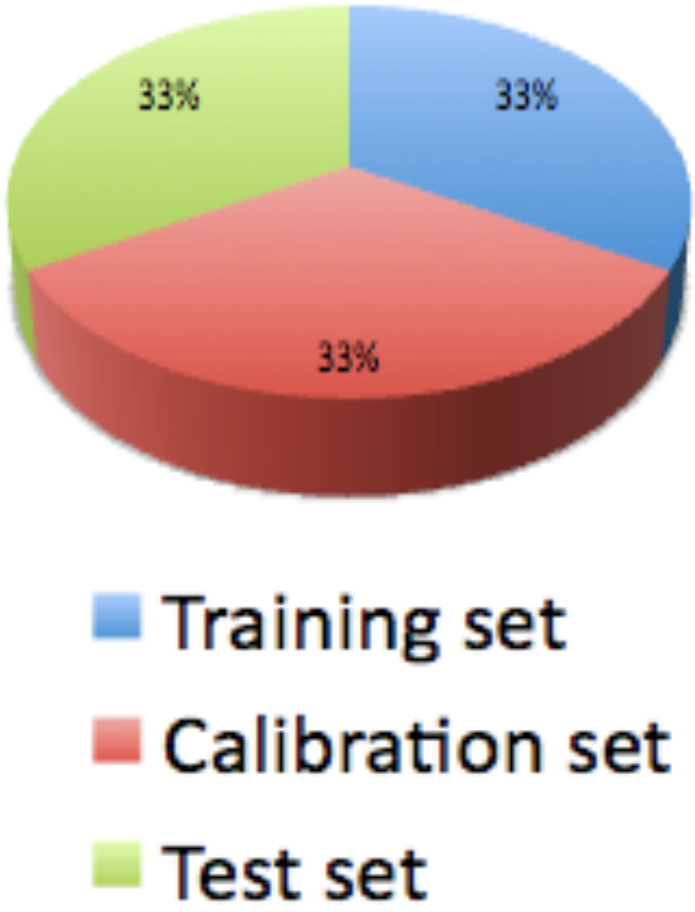
Temperature: 7 280 stations

Steps to follow:

- Data collection and preparation
- Geo-referencing of collecting locations
- Initial data exploration
- Pre-processing of dataset
- Choose modeling method
- Calibration of model
- Validation of model
- Validation of prediction results

Data for the Trait Mining model

- **Training set**
  - For the initial calibration or training step.
- **Calibration set**
  - Further calibration, tuning step
  - Often cross-validation on the training set is used to reduce the consumption of raw data.
- **Test set**
  - For the model validation or goodness of fit testing.
  - External data, not used in the model calibration.



Choosing modeling method

"There is no silver bullet" (Peterson)

- Different modeling methods or algorithms tackle issues of data quality differently.
- Some modeling methods are more sensitive to some types of data quality issues and less sensitive to other issues.
- Choosing the appropriate data modeling method depend on the types of data quality issues you discover in your respective data set.
- Identifying the appropriate method is often the process of validating performance (on an independent test set).
- It is a bad strategy to simply choose the same method that performed well in your previous studies.
- Ensembles can combine the qualities of multiple algorithms.

Data Modeling methods

- Parallel Factor Analysis (PARAFAC) (Multi-way)
- Multi-linear Partial Least Squares (**N-PLS**) (Multi-way)
- Soft Independent Modeling of Class Analogy (**SIMCA**)
- k-Nearest Neighbor (kNN)
- Partial Least Squares Discriminant Analysis (PLS-DA)
- Linear Discriminant Analysis (LDA)
- Principal component logistic regression (PCLR)
- Generalized Partial Least Squares (GPLS)
- Random Forests (**RF**)
- Neural Networks (NN)
- Support Vector Machines (SVM)
- Boosted Regression Trees (BRT)
- Multivariate Regression Trees (MRT)
- Bayesian Regression Trees

Modeling methods used by Endresen (2010), Endresen et al (2011, 2012), and Bari et al (2012).

Validation of results

- Pearson product-moment correlation (R)
- Coefficient of determination (R^2)
- Cohen's Kappa (K)
- Proportion observed agreement (PO)
- Proportion positive agreement (PA)
- **Positive predictive value (PPV)**
- **Positive diagnostic likelihood ratio (LR+)**
- Sensitivity and specificity
- Area under the curve (AUC)
  - Receiver operating characteristics (ROC)
- Root mean square error (RMSE)
  - RMSE of calibration (RMSEC)
  - RMSE of cross-validation (RMSECV)
  - RMSE of prediction (RMSEP)
- Predicted residual sum of squares (PRESS)

Classification performance

- Positive predictive value (PPV)
  - PPV = True positives / (True positives + False positives)
  - Classification performance for the identification of resistant samples (positives)
- Positive diagnostic likelihood ratio (LR+)
  - LR+ = sensitivity / (1 – specificity)
  - Less sensitive to prevalence than PPV

Some recent studies:

- Heuristic approach:
  - Sunn pest on wheat (Bouhssini *et al*. 2009)
  - Powdery mildew, *Pm3* (Bhullar *et al*. 2009)
  - Russian wheat aphid (Bouhssini *et al*. 2011)
- Multi-way approach:
  - Morphological traits for Nordic Barley landraces (Endresen 2010)
- Multivariate approach:
  - Net blotch on barley landraces (Endresen *et al*. 2011)
  - Stem rust on wheat landraces (Endresen *et al*. 2011, Bari *et al*. 2012)
  - Ug99 stem rust on wheat (Endresen *et al*. 2012)
  - Faba bean drought tolerance (Khazaei *et al*. 2013a, 2013b)
- Crop wild relatives

Salix Accessions at Alnarp, 2011 by Dag Endresen, CC-By

Sunn pest on wheat

(Heuristic FIGS approach, 2009)

- No previous sources of Sunn pest resistance had been found in hexaploid wheat.
  - 2 000 accessions were screened at ICARDA without results during 2000 to 2006.

- A FIGS set of 534 accessions was developed and screened in 2007 and 2008.
  - Starting with 16 000 wheat landraces from VIR, ICARDA and AWCC.
  - Excluding origin CHN, PAK, IND - were Sunn pest was only recently reported (6 328 accessions).
  - One accession per collecting site (2 830 accessions).
  - Excluding dry environments below 280 mm/year.
  - Excluding sites of low winter temperature below 10 degrees Celsius (1 502 accessions).
  - Reduced to 534 accessions, using PCA clustering.

10 new resistant accessions were found!

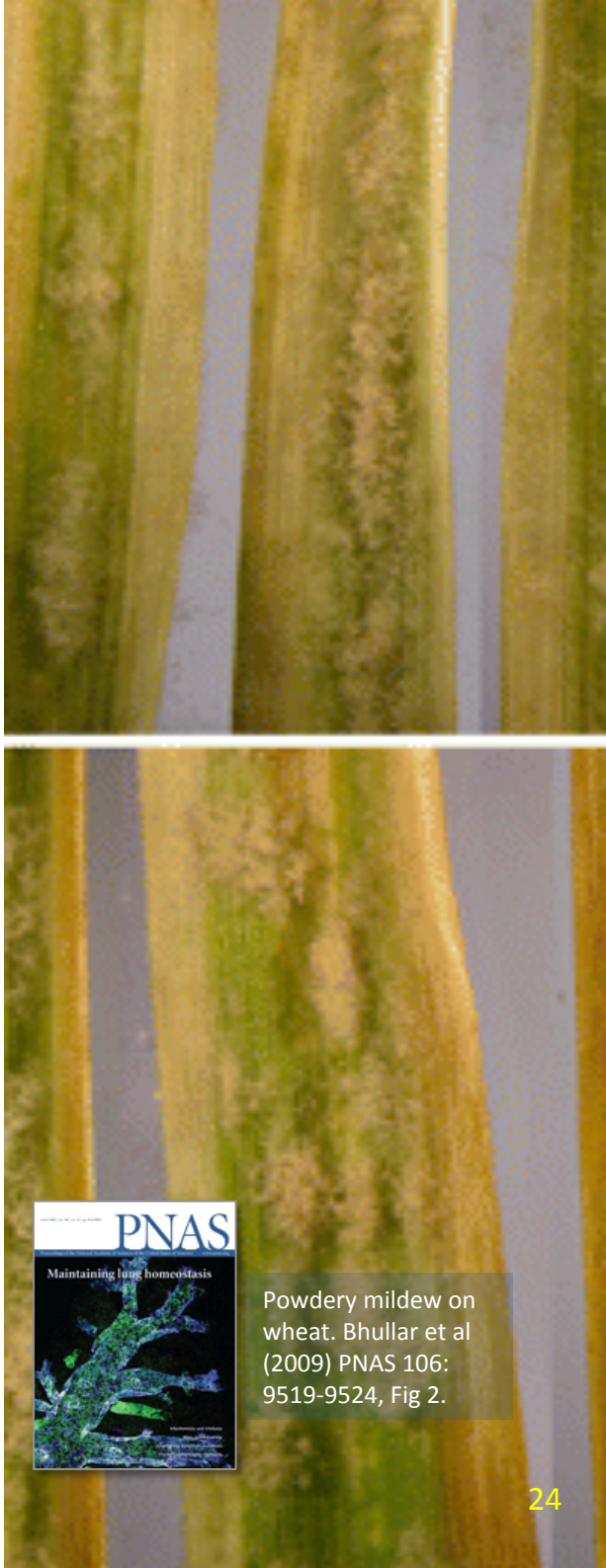
Bouhssini, M., K. Street, A. Joubi, Z. Ibrahim, and F. Rihawi (2009). Sources of wheat resistance to Sunn pest, *Eurygaster integriceps* Puton, in Syria. Genetic Resources and Crop Evolution 56:1065-1069. DOI: 10.1007/s10722-009-9427-1

Based on a slide by Ken Street, ICARDA

Powdery mildew on wheat

(Heuristic FIGS approach, 2009)

- A powdery mildew resistance set was derived based on the environmental conditions for PM hotspots.
  - Starting with 16,089 wheat landraces (6159 sites).
  - FIGS subset of 1320 wheat accessions (420 sites).
  - **211 accessions** were scored as resistant in the field trials.
  - Only 7 resistance alleles (*Pm3a* to *Pm3g*) were previously known at the *Pm3* locus.
  - **7 new resistance alleles** (*Pm3h* to *Pm3n*) were found in this study.



Powdery mildew on wheat. Bhullar et al (2009) PNAS 106: 9519-9524, Fig 2.

Bhullar, N.K., K. Street, M. Mackay, N. Yahiaoui, and B. Keller (2009). Unlocking wheat genetic resources for the molecular identification of previously undescribed functional alleles at the *Pm3* resistance locus. PNAS 106(23):9519-9524. DOI: 10.1073/pnas.0904152106.

Morphological traits in Nordic Barley landraces (2010)

Endresen, D.T.F. (2010). Predictive association between trait data and ecogeographic data for Nordic barley landraces. Crop Science 50: 2418-2430. DOI: 10.2135/cropsci2010.03.0174

Field observations by Agnese Kolodinska Brantestam (2002 to 2003)

Multi-way N-PLS data analysis

Multi-way N-PLS results: Nordic barley landraces

Experiment Site	Year	Heading days	Ripening days	Length of plant	Harvest index	Volumetric weight	*Thousand grain weight[3]*
LVA	*2002[1]*	*n.s.*	*n.s.*	*n.s.*	*n.s.*	*****	*n.s.*
LVA	2003	***	n.s.	**	**	***	*n.s.*
NOR	2002	-	*	**	***	**	*n.s.*
NOR	2003	**	***	***	*	*	*n.s.*
SWE	2002	**	***	n.s.	**	*	*n.s.*
SWE	*2003[2]*	*n.s.*	****	*n.s.*	*n.s.*	****	*n.s.*

*** Significant at the 0.001 level (p-value)

** Significant at the 0.01 level

* Significant at the 0.05 level

n.s. Not significant (at the above levels)

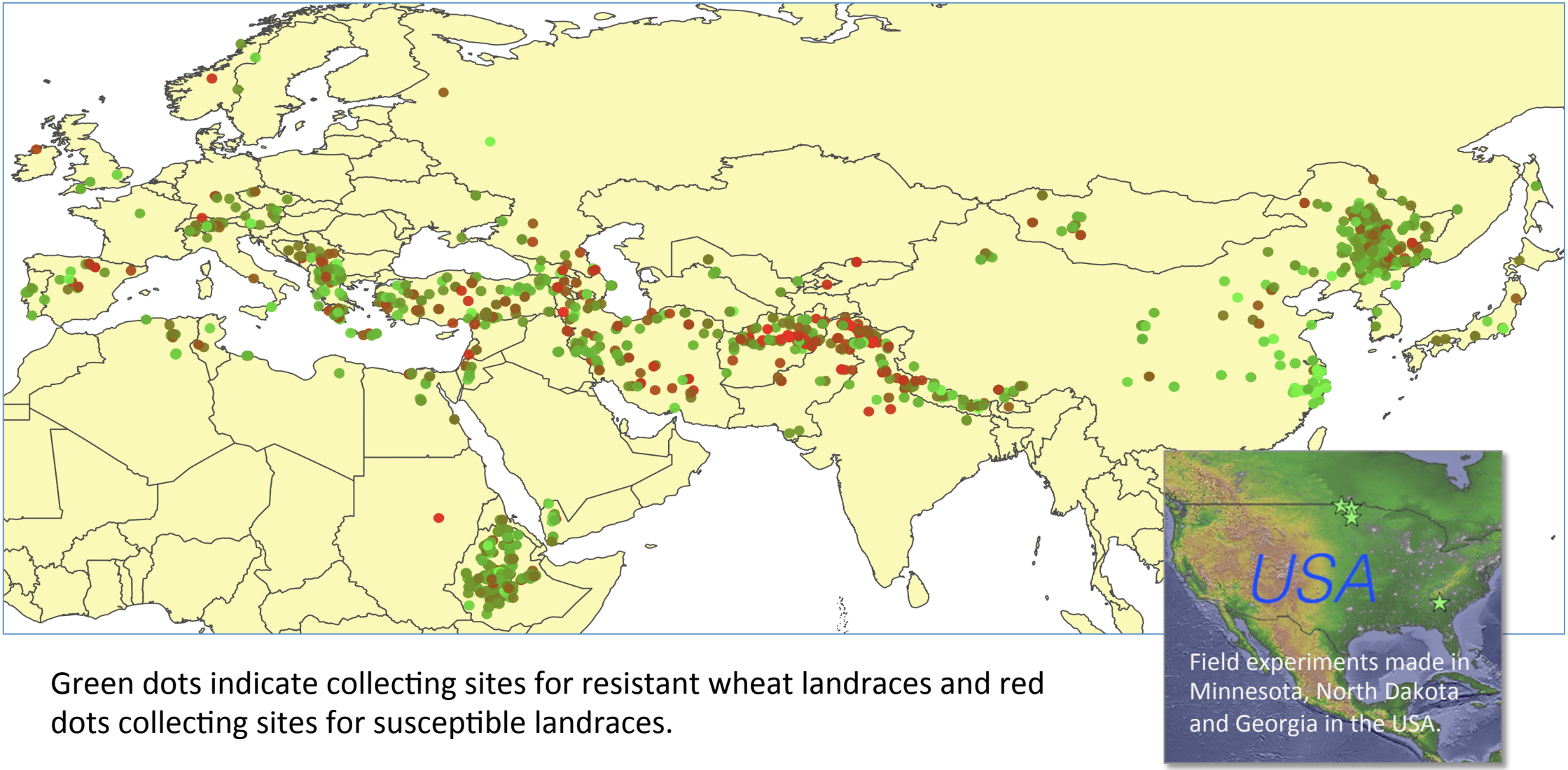
[1] LVA 2002 Germination on spikes (very wet June)

[2] SWE 2003 Incomplete grain filling (very dry June)

[3] The thousand grain weight was not predicted well.

Endresen, D.T.F. (2010). Predictive association between trait data and ecogeographic data for Nordic barley landraces. Crop Science 50: 2418-2430. DOI: 10.2135/cropsci2010.03.0174

Net blotch on barley landraces



Green dots indicate collecting sites for resistant wheat landraces and red dots collecting sites for susceptible landraces.

USDA GRIN, trait data online:
http://www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?1041

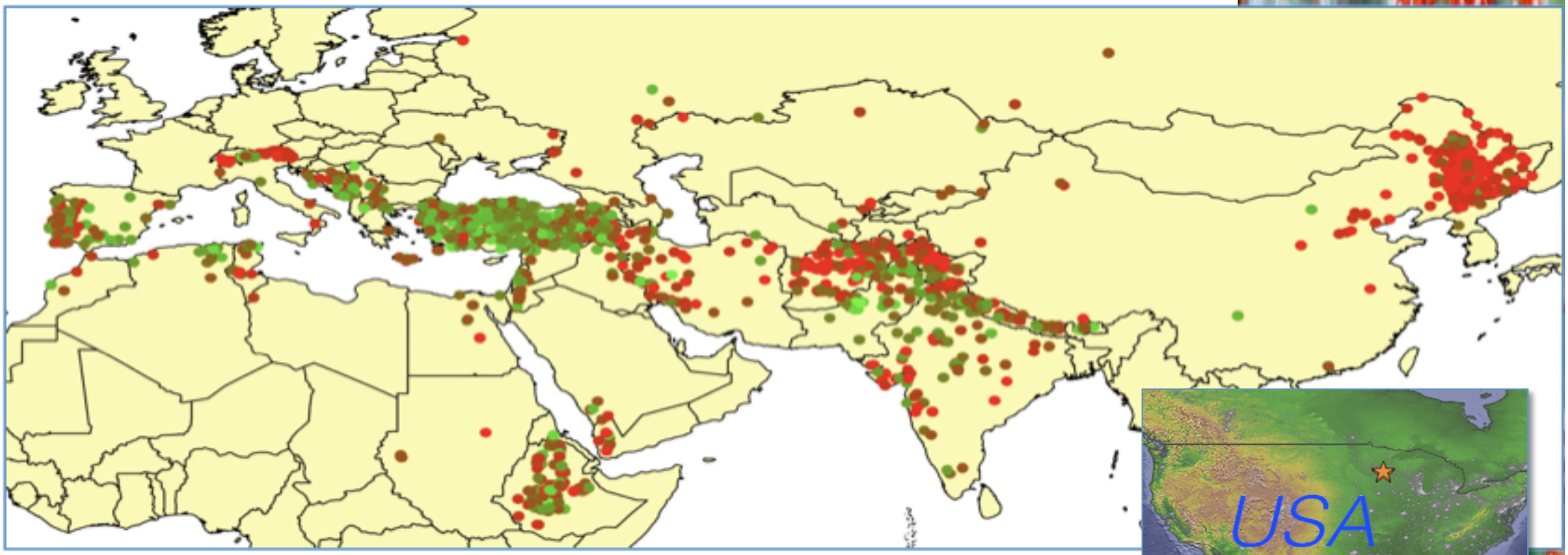
Multivariate SIMCA results: Net blotch on barley

Dataset (unit)	PPV	LR+	Estimated gain
Net blotch (accession)	**0.54** (0.48-0.60)	**1.75** (1.42-2.17)	**1.35** (1.19-1.50)
Random	**0.40** (0.35-0.45)	**0.99** (0.84-1.17)	**0.99** (0.87-1.12)
(40 % resistant samples)			

PPV = Positive Predictive Value; LR+ = Positive Diagnostic Likelihood Ratio

Endresen, D.T.F., K. Street, M. Mackay, A. Bari, E. De Pauw (2011). Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. *Crop Science* 51: 2036-2055. DOI: 10.2135/cropsci2010.12.0717

Stem rust on wheat landraces



Green dots indicate collecting sites for resistant wheat landraces and red dots collecting sites for susceptible landraces.

USDA trait data: www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?65049

Multivariate SIMCA results: Stem rust on wheat

SIMCA classification	PPV	LR+	Estimated gain
Stem rust (site)	**0.50** (0.40-0.60)	**4.00** (2.85-5.66)	**2.51** (2.02-2.98)
Random	**0.19** (0.13-0.26)	**0.94** (0.63-1.39)	**0.95** (0.66-1.33)
(20 % resistant samples)			

PPV = Positive Predictive Value; LR+ = Positive Diagnostic Likelihood Ratio

Endresen, D.T.F., K. Street, M. Mackay, A. Bari, E. De Pauw (2011). Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. *Crop Science* 51: 2036-2055. DOI: 10.2135/cropsci2010.12.0717

Photo: USDA ARS
Image k1192-1

Multivariate results: Stem rust on wheat

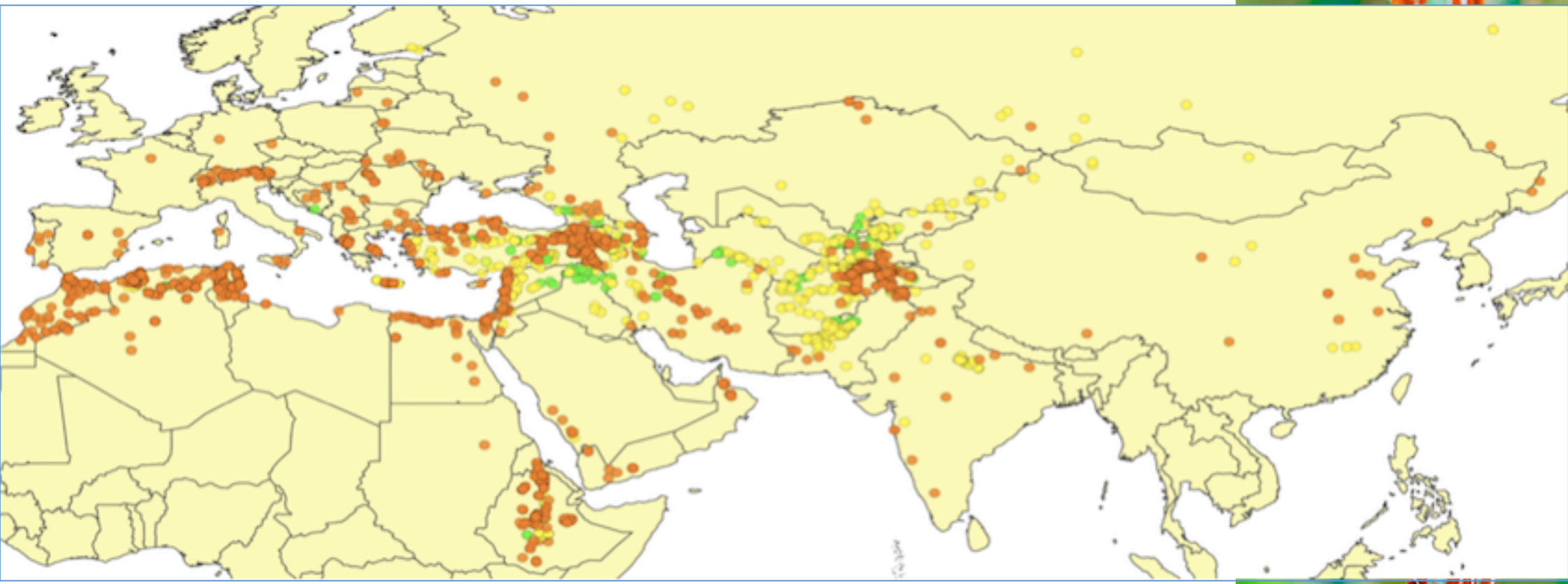
Classifier method	AUC	Cohen's Kappa
Principal Component Regression (PCR)	**0.69** (0.68-0.70)	**0.40** (0.37-0.42)
Partial Least Squares (PLS)	**0.69** (0.68-0.70)	**0.41** (0.39-0.43)
Random Forest (RF)	**0.70** (0.69-0.71)	**0.42** (0.40-0.44)
Support Vector Machines (SVM)	**0.71** (0.70-0.72)	**0.44** (0.42-0.45)
Artificial Neural Networks (ANN)	**0.71** (0.70-0.72)	**0.44** (0.42-0.46)

AUC = Area Under the ROC Curve (ROC, Receiver Operating Curve)

Bari, A., K. Street, , M. Mackay, D.T.F. Endresen, E. De Pauw, and A. Amri (2012). Focused Identification of Germplasm Strategy (FIGS) detects wheat stem rust resistance linked to environment variables. *Genetic Resources and Crop Evolution* 59(7):1465-1481. doi:10.1007/s10722-011-9775-5

Photo Ug99 stem rust on wheat, by Petr Kosina/CIMMYT

Stem rust (Ug99) on wheat



Ug99 set with 4563 wheat landraces screened for Ug99 in Yemen in 2007, with prevalence of **10.2 % resistant** accessions. True trait scores were reported for 20% of the accessions (825 samples) as training set. We used SIMCA to select 500 accessions more likely to be resistant from the remaining 3728 accessions (with the true scores hidden to the person making the analysis). This set of 500 accessions held **25.8 % resistant** samples and thus **2.3 times higher** than expected by chance.

Endresen, D.T.F., K. Street, M. Mackay, A. Bari, E. De Pauw, K. Nazari, and A. Yahyaoui (2012). Sources of Resistance to Stem Rust (Ug99) in Bread Wheat and Durum Wheat Identified Using Focused Identification of Germplasm Strategy (FIGS). *Crop Science* 52(2):764-773. doi: 10.2135/cropsci2011.08.0427

Multivariate results: Stem rust (Ug99) on wheat

Classifier method	PPV	LR+	Estimated gain
kNN (pre-study)	**0.29** (0.13-0.53)	**5.61** (2.21-14.28)	**4.14** (1.86-7.57)
SIMCA	**0.28** (0.14-0.48)	**5.26** (2.51-11.01)	**4.00** (2.00-6.86)
Ensemble classifier	**0.33** (0.12-0.65)	**8.09** (2.23-29.42)	**6.47** (2.05-11.06)
Random	**0.06** (0.01-0.27)	**0.95** (0.13-6.73)	**0.97** (0.16-4.35)
(pre-study, 550 + 275 accessions)			
Ensemble classifier	**0.26** (0.22-0.30)	**2.78** (2.34-3.31)	**2.32** (2.00-2.68)
Random	**0.11** (0.09-0.15)	**1.02** (0.77-1.36)	**0.95** (0.77-1.32)
(blind study, 825 + 3738 accessions)			

PPV = Positive Predictive Value; LR+ = Positive Diagnostic Likelihood Ratio

Endresen, D.T.F., K. Street, M. Mackay, A. Bari, E. De Pauw, K. Nazari, and A. Yahyaoui (2012). Sources of Resistance to Stem Rust (Ug99) in Bread Wheat and Durum Wheat Identified Using Focused Identification of Germplasm Strategy (FIGS). *Crop Science* 52(2):764-773. doi: 10.2135/cropsci2011.08.0427

Ug99 rust resistance trials on wheat at Njoro, Kenya. By Petr Kosina/ CIMMYT, CC-by-sa-nc.

1st Nordic Oikos symposium
& 32rd Oikos meeting
3rd to 6th February 2014
Swedish Museum of Natural History
Stockholm University

Dag Endresen
GBIF-Norway (gbif.no)
Geo-ecology research group (GEco)
Natural History Museum at the
University of Oslo (NHM-UiO)
dag.endresen@nhm.uio.no
dag.endresen@gmail.com

Michael Mackay
FIGS coordinator

Ken Street
FIGS project leader

Abdallah Bari
Scientist ICARDA

Ahmed Amri
Scientist ICARDA

Dag Endresen
GBIF-Norway