FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

PhD thesis

Dag Terje Filip Endresen

Utilization of Plant Genetic Resources

A Lifeboat to the Gene Pool

Academic advisor: Dvora-Laiô Wulfsohn and Brian Grout

Submitted: 9 February 2011

FACULTY OF LIFE SCIENCES
UNIVERSITY OF COPENHAGEN

PhD thesis

Dag Terje Filip Endresen

Utilization of Plant Genetic Resources

A Lifeboat to the Gene Pool

Academic advisor: Dvora-Laiô Wulfsohn and Brian Grout

Submitted: 9 February 2011

PhD thesis

Utilization of Plant Genetic Resources
A Lifeboat to the Gene Pool

Dag Terje Filip Endresen
Nordic Genetic Resource Center (NordGen)

University of Copenhagen
Faculty for Life Sciences
Department of Agriculture and Ecology

Academic supervisor: Dvora-Laiô Wulfsohn and Brian Grout
Assessment committee: Theo van Hintum, Nigel Maxted and Åsmund Rinnan

Submitted: 9 February 2011
Dissertation: 31 March 2011

ISBN: 978-91-628-8268-6

© Dag Terje Filip Endresen, dag.endresen@gmail.com

Printed at Media-Tryck, Lund University Press, April 2011

PDF version available online at: http://goo.gl/pYa9x
MATLAB source code available at: http://goo.gl/i52HL
Creative-Commons-Attribution (CC-By-1.0)

Cover photo: Nordic wheat landraces at Alnarp, 3 August 2010 by Dag Endresen
Available at: http://www.flickr.com/photos/dag_endresen/4998314457/

Citation: Endresen, D.T.F. (2011). Utilization of Plant Genetic Resources: A Lifeboat to the Gene Pool. PhD Thesis, Department of Agriculture and Ecology, Faculty of Life Sciences, Copenhagen University, Denmark. ISBN: 978-91-628-8268-6.

Summary

The collections of crop genetic resources are a valuable source of new genetic variation for economically important traits. New sources of useful crop traits are often identified through the evaluation of germplasm in field trials, some of which may require special treatments such as inoculation for screening purposes. The number of relevant accessions in genebank collections available to be evaluated for a specific trait is often substantially larger than the capacity or resources of the evaluation project. Thus, finding the genebank accessions most likely to posses the desired trait can be compared to searching for a needle in a haystack. The Focused Identification of Germplasm Strategy (FIGS) is a new approach used to select subsets of germplasm from genetic resource collections in such a way as to maximize the likelihood of capturing a specific trait at a higher frequency than if the subset had been selected at random. The FIGS strategy uses a range of methods to link the expression of a specific crop trait with the ecogeographic parameters of the original collection site.

In this thesis, trait mining techniques were applied in a FIGS framework. The results show that the climate layers from freely available ecogeographic databases are well suited to model and predict the reaction in these crops to biotic stress and other economically valuable traits. This result has the potential to improve the efficiency of field screening trials to find novel sources of economically valuable crop traits. This thesis provides the first multivariate data models for trait mining with FIGS. Previous published works were based on a heuristic model incorporating expert knowledge and experience when selecting a target subset. The multivariate data analysis approach allows for the evaluation of statistical significance and the degree of probability associated with a trait mining subset. Trait mining using FIGS was found to be an efficient approach for the identification of agronomical traits associated with yield and ontogeny (growth stage). This approach was also found useful for the identification of resistance to crop pathogens in germplasm from genebank collections.

Another major constraint for the use of germplasm from the genebank collections is the availability of relevant information to describe these genetic resources. New data sharing and publishing tools compatible with biodiversity informatics standards have potential for improving interoperable access to genebank datasets. The Darwin Core extension for genebanks presented with this thesis contributes to make these new data publishing tools available to the genebank community.

Keywords: Plant Genetic Resources (PGR), landraces, barley, wheat, genebank documentation, Darwin Core extension for genebanks (DwC germplasm), biodiversity informatics, Focused Identification of Germplasm Strategy (FIGS), trait mining, multiway analysis, Multilinear Partial Least Squares (N-PLS), Soft Independent Modeling of Class Analogies (SIMCA)

Preface

This thesis was prepared at the Department of Agriculture and Ecology, Faculty of Life Sciences, University of Copenhagen in partial fulfillment of the requirements for acquiring the PhD degree in agricultural sciences. The research was made in close cooperation with the Nordic Genetic Resource Center (NordGen) located at Alnarp in the south of Sweden, and also in cooperation with Bioversity International based in Maccarese outside Rome in Italy. The work presented here are the result of independent research conducted by the author of this thesis.

The topic of this thesis is the documentation of plant genetic resources and the information requirements for the rational use of the germplasm material conserved by genebank collections. The primary focus is on the Focused Identification of Germplasm Strategy (FIGS) as a new approach to select subsets of genebank accessions in such a way as to maximize the likelihood of capturing a specific trait. Another central topic of the thesis is the data exchange mechanism for the germplasm datasets required for rational utilization of the material in the genebank collections, and in particular the information required for performing new FIGS experiments.

The thesis consists of a summary report, a collection of 4 research papers, and 7 abstracts presented at international scientific conferences and seminars.

Lund, 9 February 2011

Dag Terje Filip Endresen

Acknowledgements

Many thanks to my wife Silvia who gave me the time and support required for completing a PhD. Silvia also provided valuable assistance with proofreading of the thesis manuscript. Thanks to my daughter Maria for her good heart and for listening attentive to my explanations of theories and methods of data analysis. She is now three years old and soon speaking four human languages (approximately two or three as her mother-tongue) - and with a good introduction to several computer languages. Thanks also to my own parents, Terje and Dordi and the values you gave me. I am happy and proud that my father lived to see me start my PhD research, but very sad that he did not live to see me bring it to this thesis. Thanks to my two younger brothers Pål and Ola, growing up together was good.

The most important peer to give me the gentle push to start this journey of the mind [PhD] was Sir Bent Skovmand. I am very sad that Bent did not live to see this final thesis document. Thanks to Bents family Eugenia, Astrid and Francisco for making me feel so welcome at the house of Bent in Kävlinge.

Thanks to my PhD supervisor Dvora-Laiô Wulfsohn for patience and for the gentle pushes back to the core topics when my curiosity led me a little too far into alternative research areas. Dvora-Laiô provided very careful proofreading and valuable guidance with the structure and contents of the thesis manuscript. Thanks to my co-supervisor Brian Grout for assistance when Dvora-Laiô found a new and exiting career in Chile. Thanks to all the teachers who organized the PhD training courses I attended. You all helped me with pieces to the puzzle. In particular the advanced chemometric methods for multiway analysis that Rasmus Bro was teaching me was exciting, and useful. The practical guidance by Rasmus Bro with the data analysis methods for the first paper provided a key breakthrough. I would here also like to give acknowledgement to my secondary school teacher in Biology, Dag Boman, for rooting the scientific approach to biology in a young mind.

Many thanks to my colleagues at the Nordic Gene Bank, and the institute reorganized to the Nordic Genetic Resources Center during my PhD period. I am grateful to NGB/NordGen for providing a working place and the primary source of funding for the PhD research. Mårten Huldén introduced me to genebank databases. Barbara, Lars and most recently Martin maintained the server system. Peter, Johan, Magdalena and Jonas joined me in the footsteps of previous legends like Ebbe Kjellqvist, Flemming Yndgaard, Stig Blixt, Sigfus Bjarnason, Mildred Miklasson, and Mårten, with the role of keeping the documentation of Nordic genetic resources up to high standards. Thanks to Agnese Kolodinska Brantestam for support with the research data for the first paper (I). A special thanks is directed to Axel Diederichsen and Simon Jeppson for challenging discussions and for their interest in my research topics. Axel provided marvelous help with comments to the text from the draft chapters of this thesis. Eva Jorup-Engström provided invaluable assistance through the jungle of Swedish bureaucracy - thanks! Erling Fimland from the animal genebank based in Norway provided an enthusiastic discussion partner on data analysis methods.

My second home institute during this research was Bioversity International based in Rome. Bioversity was known as International Plant Genetic Resources Institute prior to 2007. Samy Gaiji, Milko Škofič and Rajesh Sood became good friends and shared with me from their valuable experience on the documentation of plant genetic resources. During the recent years Michael Mackay and Elizabeth Arnaud (at Bioversity) provided valuable support. A special thanks to Michael Mackay (Bioversity) and Kenneth Street based at ICARDA in Syria for collaboration and help with the Focused Identification of Germplasm Strategy (FIGS).

From the genebank documentation community I wish to acknowledge the European Cooperative Programme for Plant Genetic Resources (ECPGR) Documentation and Information Network coordinating group. Our meetings and not least the challenges we faced together during the development of the EURISCO (European Catalogue of *ex situ* genebank collections) platform was a good place to learn genebank documentation. A very special thanks to Helmut Knüpffer, head of documentation at the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) in Gatersleben (Germany) and my co-author for the third paper (III). Helmut provided careful proofreading of chapter 5 and valuable feedback on nomenclature and ecogeography.

The secretariat staff at the Global Biodiversity Information Facility (GBIF) and the members of the GBIF and Biodiversity Information Standards (TDWG) network provided a valuable forum for exchange of ideas on biodiversity informatics. In particular my colleagues in the GBIF Nodes committee became not only close friends, but also valuable discussion partners. Donald, Markus, Tim, Samy, José, Kyle, David, Éamonn, Vishwas, Lawrence, Juan, Paco, Steve, Mélianie, Mihail, Isabel, Cees, Guy, Hannu, Walter are only some of the many names from the GBIF *'family'*.

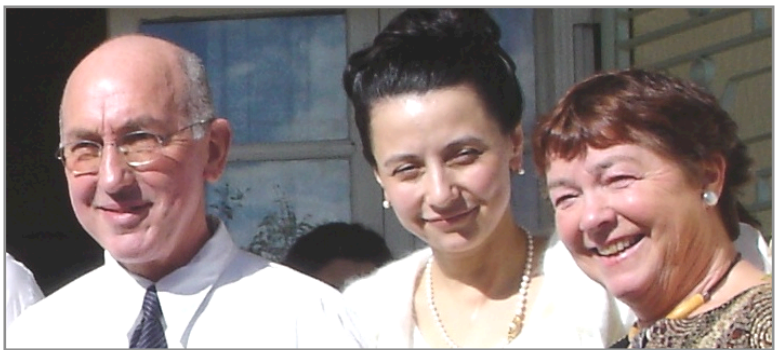
I am very proud that two of the most influential personalities and capacities from the European genebank community, from respectively Wageningen and Birmingham, was part of the assessment committee for the thesis defence. Of equal importance is the contribution to the assessment committee from one of my teachers on the advanced chemometric methods from Copenhagen University. Drs. Åsmund Rinnan, Theo van Hintum and Nigel Maxted made very useful corrections and feedback that are included in this final version of the thesis.

Dedication

For my father Terje Endresen (1941-2009) and my professor Bent Skovmand (1945-2007). You were both a great inspiration and provided each in your own way the essential sparkle for starting this research project. For my family Maria and Silvia, father, mother and two brothers, Ola and Pål, for providing sense to it all [life].

Bent Skovmand at Svalbard, June 2006. Photo by Lene Krøl Christensen.

Terje Endresen in Arendal, Norway, March 2000.



Terje Endresen, Silvia Filip-Endresen, and Dordi Endresen in Tunis, Tunisia, November 2004.

Ola and Pål at Torungen outside Arendal, Norway, June 2006.

Maria and Silvia in Lund, Sweden, August 2010.

Papers included in the thesis

The thesis is based on the following three manuscripts, referred to by their respective Roman numerical. ***Paper I*** *is submitted to Biodiversity Informatics (ISSN: 1546-9735).* ***Paper II*** *is reprinted with the kind permission from the publisher (Crop Science, ISSN: 1435-0653).* ***Paper III*** *is submitted to Crop Science (ISSN: 1435-0653) and* ***Paper IV*** *is presented as a draft manuscript.*

PAPER I:

Endresen, Dag Terje Filip and Helmut Knüpffer. The Darwin Core extension for genebanks opens up new opportunities for sharing genebank datasets. *Submitted to Biodiversity Informatics on 31 Jan 2011.*

PAPER II:

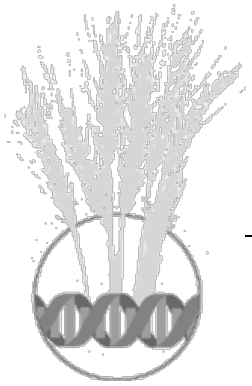
Endresen, Dag Terje Filip (2010). Predictive association between trait data and ecogeographic data for barley landraces. Crop Science 50(6): 2418-2430. DOI: 10.2135/cropsci2010.03.0174. Available as open access at https://www.crops.org/publications/cs/articles/50/6/2418

PAPER III:

Endresen, Dag Terje Filip, Kenneth Street, Michael Mackay, Abdallah Bari, and Eddy De Pauw (2011). Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. *Submitted to Crop Science on 21 December 2010. Conditionally accepted 6 Feb 2011; first revision submitted 9 March 2011; accepted on 10 April 2011.*

PAPER IV:

Endresen, Dag Terje Filip, Kenneth Street, Michael Mackay, Abdallah Bari, and Eddy De Pauw (*draft manuscript*). Sources of resistance in bread wheat to stem rust (Ug99) identified using the Focused Identification of Germplasm Strategy (FIGS). *This draft manuscript constitutes my contribution to a collaborative paper with a follow-up study to PAPER III by the same authors.*



Abstracts included in the thesis

ABSTRACT 1:
Endresen, D.T.F. and B. Skovmand (2006). Trait mining in gene banks. American Society of Agriculture, ASA-CSSA- SSSA, Indianapolis, USA, 12-16 November 2006. Available at http://a-c-s.confex.com/crops/2006am/techprogram/P26713.HTM

ABSTRACT 2: *Peer review*
Endresen, D.T.F., J. Bäckman, H. Knüpffer, and S. Gaiji (2006). Exchange of germplasm dataset with PyWrapper/BioCASE. p. 8. *In*: Belbin, L., A. Rissoné, and A. Weitzman (eds). Proceedings of TDWG 2006. 16 October 2006, Missouri Botanical Garden in St. Louis, Missouri, USA. ISBN: 1-930723-56-3. Available at http://www.tdwg.org/proceedings/article/view/64

ABSTRACT 3: *Peer review*
Knüpffer, H., D.T.F. Endresen, and S. Gaiji (2007). Integrating genebanks into biodiversity information networks. p. 34-35. *In*: Hauptvogel, P., D. Benediková, and R. Hauptvogel (eds). 18th EUCARPIA Conference. Genetic Resources Section. Plant Genetic Resources and their Exploitation in the Plant Breeding for Food and Agriculture. 23 May 2007, Piešťany, Slovak Republic. ISBN: 978-80-88872-63-4. Available at http://www.cvrv.sk/fileadmin/CVRV/novinky/book_of_abstracts.pdf

ABSTRACT 4:
Endresen, D.T. (2008). Biodiversity data exchange software, hands-on exercises with TAPIR software. 3 Dec 2008, Stockholm Biodiversity Informatics Symposium 2008 (SBIS2008), Swedish Museum of Natural History, Stockholm, Sweden. Available at http://artedi.nrm.se/sbis2008/

ABSTRACT 5: *Peer review*
Endresen, D., S. Gaiji, and T. Robertson (2009). Darwin Core germplasm extension and deployment in the GBIF infrastructure. p. 78. *In*: Weitzman, A.L. (ed). Proceedings of TDWG 2009. 12 Nov 2009, Montpellier, France. Available at http://www.tdwg.org/proceedings/article/view/464

ABSTRACT 6:
Endresen, D. (2010). A Lifeboat to the Gene Pool. Predictive association between trait data and ecogeographic data for identification of trait properties useful for improvement of food crops. Vavilov and Waterman seminar. 12 May 2010, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany.

ABSTRACT 7: *Peer review*
Gaiji, S., D. Endresen, J. Nordling, S. Dias, E. Arnaud (2010). Beyond Darwin Core: Challenges in mobilizing richer content. p. 15-16. *In*: Weitzman, A.L. (ed). Proceedings of TDWG 2010. 28 Sep 2010, Woods Hole, Massachusetts, USA. Available at http://www.tdwg.org/conference2010/

Contents

List of Figures

List of Tables

CHAPTER 1

1. Introduction

This thesis focuses on the use of the genetic diversity conserved by germplasm genebanks and more specifically some of the major obstacles to their rational and efficient use in crop improvement and research. The lack of complete and readily available information on genebank accessions causes problems for germplasm users to identify the genebank samples likely to hold the genetic diversity they search for. The lack of descriptive information including the so-called characterization and evaluation data was identified as a major obstacle to the identification of relevant samples for crop improvement and other relevant uses (Brown et al., 1989; FAO 1997).

The so-called core sampling strategies (chapter 6; Frankel, 1984; Frankel and Brown, 1984a, 1984b) achieved great popularity among genebank managers, but was less useful for the users of the genebank collections - the plant breeders and crop researchers. The genebank curators, strives to maintain plant genetic diversity in *ex situ* collections (Rao et al., 2006). Their objectives include avoiding genetic drift caused by storage conditions and to maintain the maximum genetic diversity from the original samples for future uses. Plant breeders and most crop researchers are more often interested in a small subset of the genetic diversity from the large genebank collections. They are generally interested in a target trait property for a specific breeding purpose or research hypothesis (Mackay, 1986, 1990). Uncritical base broadening of the breeding material is uncommon, while the search for economically useful trait properties for crop improvement is more common (*cf*. Acquaah, 2007).

This thesis strives to make a contribution on bringing the available information on germplasm samples more readily available (1) and to provide an approach to use the limited information available in new ways to better meet the needs of germplasm users (2). The objective of more readily available genebank information is addressed through the use of information sharing technologies and web services (PAPER I). The second objective on new ways to use the available information was tackled by the utilization of Focused Identification of Germplasm Strategy (FIGS; Mackay and Street, 2004). This thesis provides an implementation of the FIGS approach using predictive multivariate data analysis to *'fill in the gaps'* in response to the general lack of descriptive evaluation data (PAPER II, III, IV).

1.1 Objectives and research hypothesis

The utilization of plant genetic resources is limited by the access to relevant information and by efficient data analysis and sampling strategies designed for the needs of the genebank users such as plant breeders searching for a target trait (FAO, 1997, 2010).

The main objectives of this research were to:
(1) Provide new solutions for data exchange and the sharing of genebank data.
(2) Provide an implementation for trait mining with FIGS using predictive multivariate computer models.

The research hypothesis can be stated as: (1) Improved access to genebank data with the use of modern data sharing technologies will improve the accessibility and usefulness of this material for the users. (2) Trait mining in genebank collections using the FIGS approach will assist germplasm users to identify useful genebank samples representing the target genetic diversity

they need for crop improvement and research questions. (3) There is a link between ecogeography at the collecting site and trait data for landraces that can be exploited by trait mining in a FIGS framework.

1.2 Assumptions of trait mining

Behind the adoption of the computational approaches used in the thesis, lie a number of assumptions as follows: (1) Trait properties in germplasm samples are linked to the ecogeography at the original collecting site. (2) Predictive computer models can exploit this link to select germplasm samples with a higher concentration of a target trait property. (3) Multivariate data analysis methods can be used to build such predictive computer models. (See figure 1.1 and 1.2)

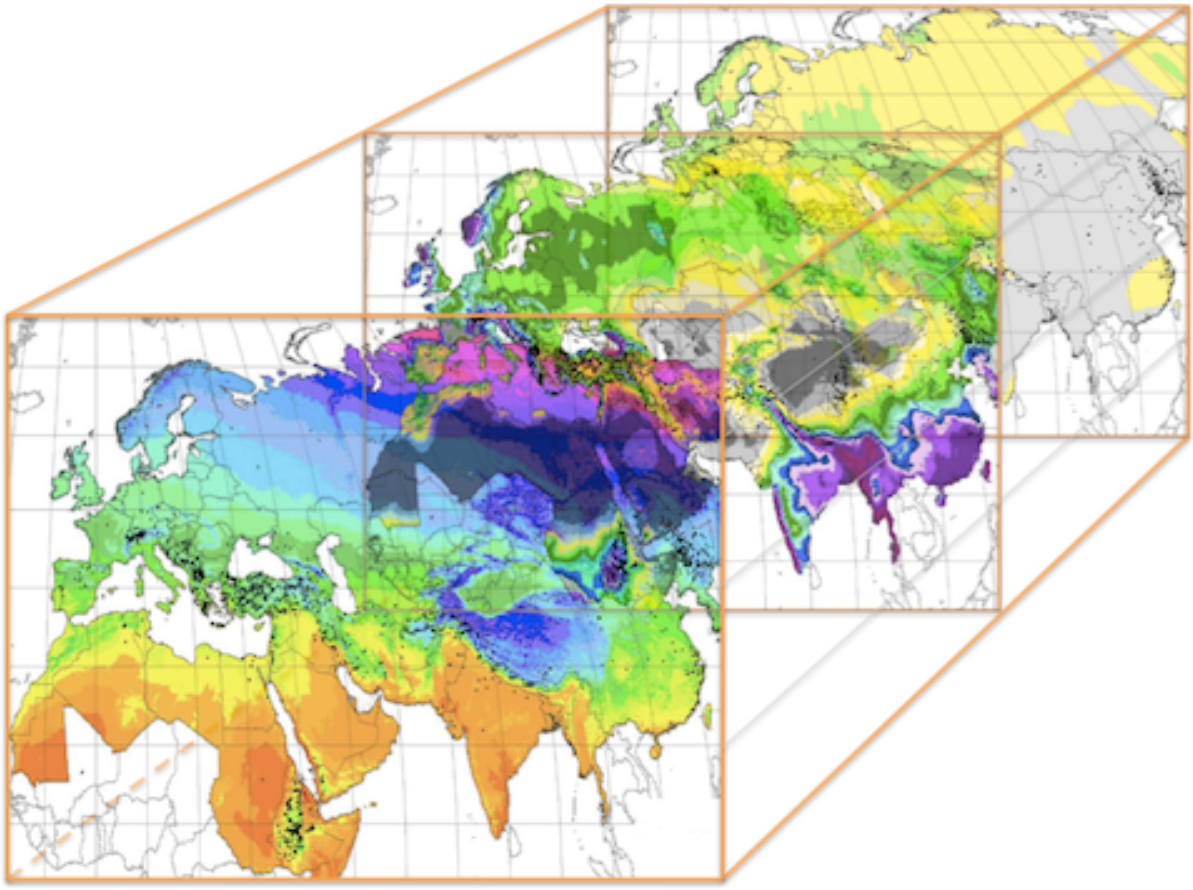


Figure 1.1: Illustration of trait mining with ecoclimatic GIS layers. GIS layers included in the illustration are from the ICARDA ecoclimatic database, average: annual temperature (front), annual precipitation (middle), and winter precipitation (back) (De Pauw, 2008)

It is possible that the farmers who participate in the development of the landraces enhanced the predictive link between ecoclimate and the crop traits. The farmer was likely to select plants that were more adapted to the ecoclimatic conditions where she lives and farms. She was also likely to introduce germplasm from other locations, including germplasm material from locations of similar ecoclimate through seed exchange. It is thus not impossible that the cultivation of plants have strengthened the link between crop traits and the ecoclimatic conditions (figure 1.3).

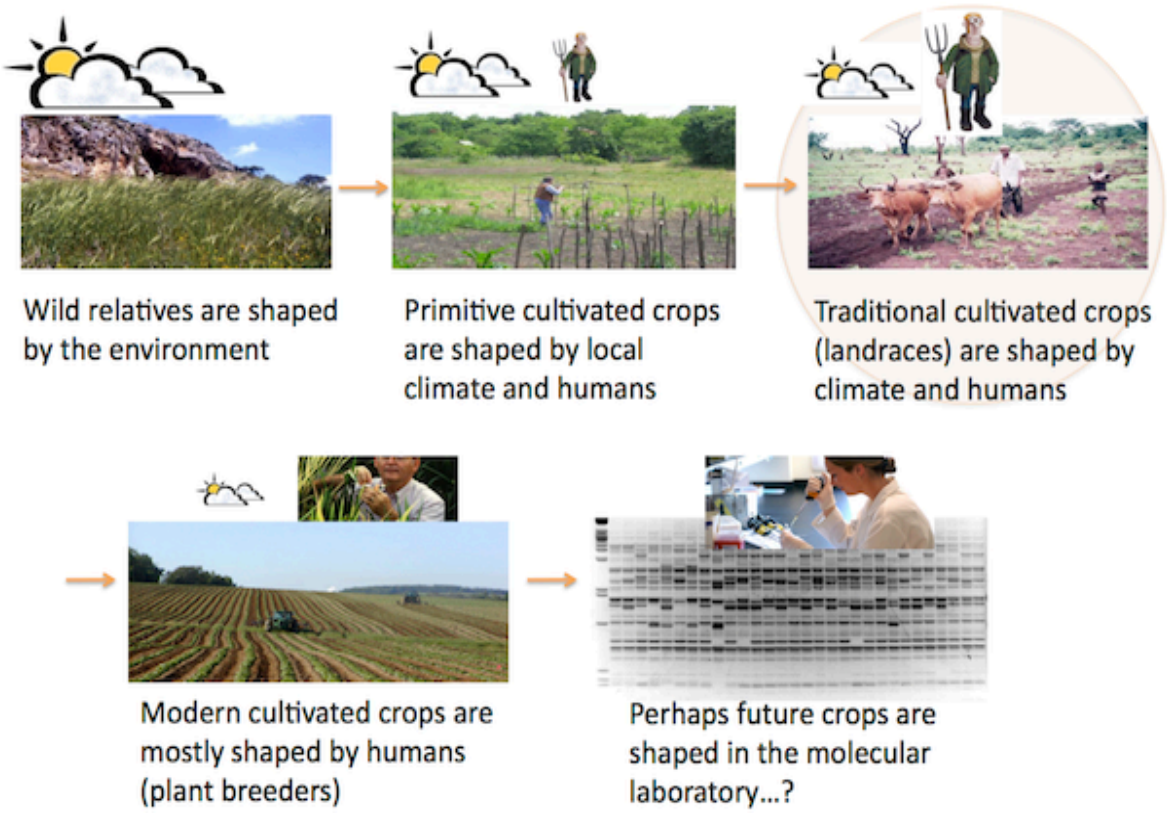


Figure 1.2: Expected adaptation to ecoclimatic factors varies from wild relatives of the cultivated plants, through the landraces to the modern cultivars.

Figure 1.3: Farmers select germplasm material more suitable for cultivation under the ecoclimatic condition that prevails where they live. The farmer can select plants or mixtures of agroecotypes that are more adapted to the conditions of where she lives, and she can introduce germplasm from areas of similar ecoclimate through seed exchange.

1.3 Introduction to the chapters

This first chapter (1) introduces the thesis research objectives and hypothesis. The second chapter (2) provides an historic background to plant genetic resources, the early utilization and first conservation in *ex situ* genebanks. The third chapter (3) covers why we need to conserve and use genetic resources. Documentation of contents and properties of genebank collections are vital for their usefulness, and covered by the forth chapter (4). Ecogeographic analysis including names and georeferencing are required preparations before proceeding with trait mining experiments and described by the fifth chapter (5). Focused Identification of Germplasm Strategy (FIGS) was initially proposed as a complement to the Core collection sampling strategies. Core collections and FIGS are presented with the sixth chapter (6). The seventh chapter (7) presents the results from this PhD project, and the eight chapter (8) proposes some directions for future work based on these results.

Chapter 2: Introduction to plant genetic resources
The history of cultivated plants starts with the domestication of the first plants more than 10 000 years ago. The diversity of cultivated forms that is available today is the result from millennia of cultivation and selection of the most interesting forms. In modern times the botanical gardens established in Europe provided a useful service for botanical classification and other research on plants, including the cultivated plants. The botanical gardens contributed to the introduction of useful plants and to the exchange of these plants with their *index seminum* [list of seeds, seed catalog]. So-called modern plant breeding and the systematic development of cultivars for improved yield, was started during the 19th century with a major *"boost"* in 1900 after the *'re-discovery'* of Mendel's work on heredity. However the rapid development of improved forms created a controversy regarding the ownership of these genetic resources and the fair sharing of benefits arising from their exploitation. The attempts to develop international legislation to address this controversy led to a series of parallel initiatives that still call for closer integration: UPOV, WTO, FAO, and UNEP are the most relevant organizations. The Vavilov Institute located in St Petersburg is recognized as the first modern genebank. In particular the activities after Nikolai Vavilov was appointed director in 1920 laid down the foundation for modern genebanks. This thesis addresses the use of the plant genetic resources preserved and made available from a worldwide network of genebanks; and in particular the identification of useful genetic diversity using the *Focused Identification of Germplasm Strategy* (FIGS) in landraces that are the result from millennia of farming.

Chapter 3: The use of plant genetic resources
Why do we need plant genebanks in the first place? What are the most important challenges of mankind today, and how do genebanks contribute to address these challenges? What is the germplasm conserved and made available by genebanks used for? What do the plant breeders and crop scientists requesting germplasm material from the Nordic genebank report that they plan to use the materials to achieve? How can the genebanks contribute to make their collections more useful? Chapter two addresses these questions. The enigma of insufficient funding for sustainable conservation of plant genetic resources for the future needs is also briefly discussed. This thesis focuses on two aspects to increase the usefulness of genebank collections. (1) The availability of accurate and relevant information on the germplasm material is crucial for the rational use. (2) The FIGS approach will assist the plant breeder and crop scientists to find more relevant accessions in genebank collections in situations when they look for a target trait property.

Chapter 4: Genebank documentation
FAO recently published the Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. This report and similar reports published earlier, point to the lack of available and relevant documentation on the germplasm material provided from genebanks as one of the most important obstacles for their rational use. Chapter 3 addresses (1) the development of genebank management software to manage and improve on the quality of the information; and (2) the development of modern data exchange mechanisms to improve the accessibility of the existing information stored in the genebank database system, including linking to relevant datasets from other sources.

Chapter 5: Ecogeographic data analysis
Biological diversity and agrodiversity evolves in an ecogeographic context. The environment is one of the most influential forces driving the evolutionary process. Survival of the fittest offspring lead to the adaptation of plants to the environment they are exposed to. The botanical classification of plants and so-called *'ecotypes'* including the cultivated plants, provides an essential platform to support experimental studies and not least, the communication between people making these experiments, or people wanting to use and exploit the plant genetic diversity in genebank collections. Ecological Niche Modeling provides a popular tool to study and predict species distributions and other interactions with the ecogeography. The approach to FIGS analysis in this thesis can be seen as a special case of ecological niche modeling, where the FIGS model describes the distribution of a target crop trait in the ecogeographic space, rather than the distribution of the species. This chapter also describes some of the previous studies linking crop trait properties and ecogeographic data including climate data.

Chapter 6: Core collections and FIGS
The *Core collection* concept was proposed by Frankel in 1984 as a strategy to improve access and rational use of genebank collections. The FIGS (Focused Identification of Germplasm Strategy) approach can be seen as a response to the Core collection strategy, and was originally proposed as an alternative core collection sampling strategy. Chapter 4 introduces the Core collection concept, and also presents some of the available software implementations for core subset sampling. The FIGS concept will also need to be ported to user-friendly software implementations to support the wider use of FIGS by plant breeders and crop scientists.

Chapter 7: Trait mining
This thesis provides some of the first practical results to support the validity and usefulness of the FIGS concept. The data analysis approach followed is described so that other scientists can follow the same procedure to make similar trait mining experiments with their own datasets. The materials and methods can also be seen as a first step towards the development of a software implementation for making FIGS analysis. This thesis introduces multi-way data analysis for genebank datasets. The results presented here indicate that multi-way methods can extract more information from genebank datasets than comparable standard multivariate methods (PAPER II).

Chapter 8: Future work
This thesis provides only the early exploration of the possibilities from trait mining with FIGS. The link between the ecoclimate and the crop traits could be studied and characterized further in a recommended follow-up study with germplasm from wild relatives of the cultivated plants in replacement of the landrace gene pools exclusively studied by the current FIGS experiments.

CHAPTER 2

2. Introduction to plant genetic resources

2.1 Biodiversity and genetic diversity

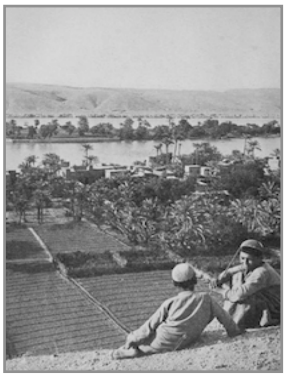
Biodiversity includes a rich diversity of species adapted to a wide range of ecological conditions. Populations of species adapted to a subset of the species habitat are called *'ecotypes'* or *'ecospecies'*. Typically ecotypes show distinct phenotypic properties in response to their specific ecological environment. Traditional crop cultivars called *'landraces'* (Zeven, 1998) can be seen as the agricultural equivalent of the ecotypes in wild species. Crop plants have evolved in response to the ecoclimatic and agricultural environment as human civilization developed and spread out across the planet. This phase of spread and adaption to new ecological conditions and new agricultural practices resulted in a very large genetic diversity represented by distinct landraces of the crop species and has steadily progressed the last 10 000 years. Modern plant breeding has reversed this steady increase in crop genetic diversity by the development of uniform modern high yielding cultivars outcompeting the diverse mosaic of previous landraces grown around the world. This alarming effect on genetic erosion in the cultivated crops was increasingly more visible resulting in the systematic collection and conservation of the previous crop genetic diversity represented by the disappearing landraces (Harlan and Martini, 1936; Bennett, 1965; Harlan, 1975; Negri et al., 2009). Under the coordination of the Food and Agriculture Organization of the United Nations (FAO) a network of international and national genebanks were established based on the model of the Russian Institute of Plant Industry located in Saint Petersburg, which was established in 1894 as the Bureau of Applied Botany. These events will be covered in more detail below.

2.2 Early domestication of plants

The beginning of agriculture and the first domestication of plants and animals are considered to have taken place in the so-called *'Fertile Crescent'* more than 10 000 years ago. The Fertile Crescent is shaped as an arc around the Syrian Desert with the Mediterranean Sea to the west, the Anatolian mountain ridges in the north, and the Zagros Mountains of present day Iran to the west. The westerns part of the Fertile Crescent is known as the Levant and covering present day Jordan, Israel, the Palestinian Territories, Lebanon, Syria, southeastern Turkey. The eastern part follows the Mesopotamian river valley of Euphrates and Tigris, in present day Iraq (Bellwood, 2005). For most of the period since the advent of the first domesticated forms the farmer selected what she considered the best plants as the seed source for sowing the next season, leading to a gradual improvement of the cultivated plants. This is how the primitive cultivars and landraces were adapted to the local conditions through long-term cultivation (Zeven, 1998). *"However some say that the people of Syria use no cultivation, except cutting out wood and watering, also that the date-palm requires spring water rather than water from the skies; and that such water is abundant in the valley in which are the palm-grooves. And they add that the Syrians say that this valley extends through Arabia to the Red Sea, and that many profess to have visited it, and that it is in the lowest part of it that the date-palms grow"* (Theophrastus, ca 300 BC, translation by Sir Arthur Hort, 1916:138).

The early domestication of the date tree (*Phoenix dactylifera* L.) played an important role in the development of the Babylonian and Assyrian civilizations (Roberts, 1929). The early scientific history of the Arabs is also described by Abū Yaḥyā, Zakariyā Ibn Muḥammad Ibn Maḥmūd, al-Qazwīnī (أبو يحيى زكريا بن محمد القزويني) (1203-1283). The evolved Babylonian and Assyrian tradition of hand-pollination of the perennial date tree seems however primarily to

have been for the production of fruit and not for the production of seeds or for the purpose of breeding the plant (Roberts, 1929). *"The date has a striking resemblance to man, through the beauty of its erect and slender figure, its division into two distinct sexes, and the property, which is peculiar to it, of being fecundated by a sort of union"* (Zakariya al-Qazwini, ca 1283 *cf.* Roberts, 1929:11).



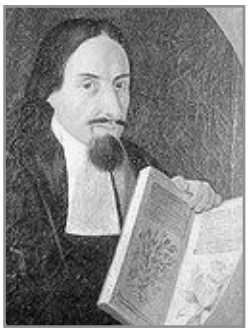
Date palm trees in Mesopotamia (from Roberts, 1929)

Assyrian priests hand-pollinate date tree (from Roberts, 1929, originals at British Museum, Nimrod Gallery)

Date palm trees in Doux, Tunisia, by Dag Endresen 21 Jan 2005.

2.3 Early collections and gardens of plant material

Luca Ghini (ca 1490-1556) designed and established the first modern botanical gardens in Europe at the University of Pisa in 1544, and at the University of Padua and University of Florence in 1545. During the following decades, similar botanical gardens were established all across Europe, based on the design of the Padua garden. The most important include Bologna (1547), Zürich (1560), Leiden (1577), Leipzig (1579), Montpellier (1598), Paris (1597), Oxford (1621), Berlin (1679), Edinburgh (1680), and Amsterdam (1682) (Stafleu, 1969).



Luca Ghini (ca 1490-1556) designed the first botanical gardens.

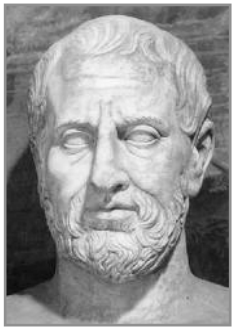
'Orto Botanico di Pisa', established in 1544, relocated in 1563, and again 1591 to the current location. Photo by Laurentius, May 2006.

'Orto Botanico di Padova' (Padua Botanical Garden, established in 1544-1545) with the Saint Anthony basilica in background (16th century).

The Ancient Greek philosophers Aristotle (*'the father of science'*, 384-323 BC) and Theophrastus (*'the father of botany'*, 371-287 BC) established the earliest known zoo and botanical garden in Europe around 330 BC with active support from Alexander the Great (356-323 BC). Alexander brought numerous animals and plants from his expeditions of conquest back to Greece, to be included in the zoo and the garden of the Aristotle Lyceum [school] (Thanos, 1994). Theophrastus reported in his classic book *'Enquiry into plants and minor works on odours and weather signs'* about botanical gardens and cultivars in ancient Babylon times. This was during the times shortly after the conquest of Babylon by Alexander the Great: *"The best kind alike in size and in quality, whether of the white or black variety, is that which in either form is called 'the royal palm'; but this, they say, is rare; it grows hardly anywhere except in the park of the ancient Bagoas* [Persian Chief Minister under King Artaxerxes III of Persia

(425-338 BC), Bagoas died 336 BC]*, near Babylon"* (*cf*. Hort, 1916:139), and: *"Thus they say that ivy and olive do not grow in Asia in the parts of Syria which are five days' journey from the sea (…) However, when Harpalus [aristocrat of Macedon] took great pains over and over again to plant it [ivy tree] in the gardens of Babylon, and made a special point of it, he failed: since it could not live like the other things introduced from Hellas"* (Theophrastus, ca 300 BC, translation by Hort, 1916:311).


Aristotle
(384-323 BC)

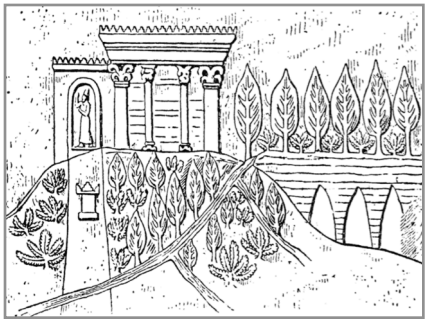

Theophrastus
(371–287 BC)


Alexander the Great
(356-323 BC)

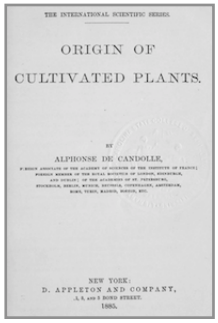

Nebukadnessar II
(605-562 BC)

Sir Arthur William Hill (1875-1941) was the assistant director of the Royal Botanic Gardens in Kew when he attended the 25th anniversary of the Missouri Botanical Garden in Saint Lois, MO, USA. Hill took later over as director of the Kew Gardens in 1922. The manuscript Hill presented for the conference proceedings describes *'The history and functions of botanical gardens'* including botanical gardens established in ancient Egypt, China, and Mexico: *"Prior to the interest displayed by the Greeks in the vegetation of the earth and quite independent of their influence we find evidence of the formation of gardens in Egypt, Assyria, China, and subsequently in Mexico - gardens not strictly botanic in our more modern sense but enclosures set apart for the cultivation of plants of some definite economic or aesthetic value* (…) *The earliest garden of which we have any representation is the Royal Garden of Thotmes III* [Thutmose III, Pharaoh of the 18th dynasty of ancient Egypt, died ca 1425 BC] *of about the year 1000 B. C., which was planned by Nekht, head gardener of the gardens attached to the Temple of Karnak. (...) The Chinese, however, should, as might be supposed, be credited with being the real founders of the idea of botanic gardens, since it is clear that collectors were dispatched to distant parts and the plants brought back were cultivated for their economic or medicinal value. The semi-mythical Emperor Shen Nung, of the twenty-eighth century B.C., is considered to be the Father of Medicine and Husbandry and is said to have tested the medical qualities of herbs and discovered medicines to cure diseases. (...) Montezuma* [Moctezuma II (1466–1520), ninth Aztec Emperor] *had extensive gardens filled with fragrant shrubs and flowers and especially with medicinal plants. (...) The gardens at Iztapalan and Chalco are said to have been stocked with trees and plants scientifically arranged"* (Hill, 1915:185-187).


Shen Nung (c 3000 BC)
'Father' of Chinese Agriculture.


The Assyrian Kuynjik relief from ca 650 BC may show hanging gardens in Babylon.


Alphonse Pyramus de Candolle (1806-1893)

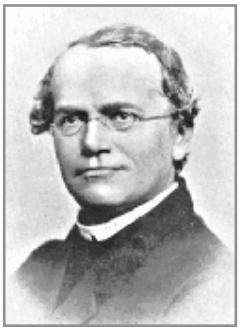


Origin of Cultivated Plants (de Candolle, 1883)

Alphonse Lois Pierre Pyramus de Candolle (1806-1893) was a French-Swiss botanist and the son of Augustin Pyramus de Candolle (1778-1841). Father and son were both famous botanists, but we will here focus on the works by de Candolle, the younger, on the origin of cultivated plants. Candolle proposed some of the first systematic theories on why people started to cultivate plants and to where this first happened (Candolle, 1883, 1884a, 1884b). He describes the diffusion of superior cultivated species and forms, and generally describes the origin of agriculture as an achievement leading to less work and more leisure time available to those people adapting this practice.

2.4 Plant breeding after Mendel discovered the principles of heredity

After the re-discovery of the work by Mendel in 1900 by Hugo de Vries and Carl Erich Correns (Mendel, 1866; Vries, 1900; Correns, 1900) the agricultural research into hybridization and new cultivars caused a rapid increase in agricultural yield (Endersby, 2007). Previously also Erich von Tschermak-Seysenegg was credited as the third to re-discover the work of Mendel (Tschermak, 1900), but a closer examination reveals that his understanding of the work by Mendel seems to be incorrect (Monaghan and Corcos, 1986, 1987).


Gregor Mendel (1822-1884)

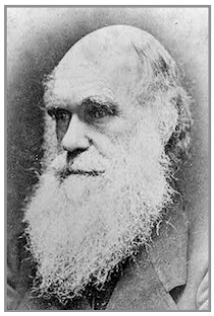

Hugo de Vries (1848-1935)

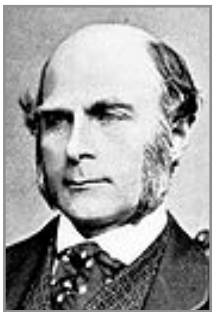

Carl Erich Correns (1864-1933)


Erich von Tschermak-Seysenegg (1871-1962)

Notice however that Mendel was not the first to study the laws of heredity - if at all this was the aim of his studies (Kampourakis, 2010). Charles Darwin presented his own theory of heredity as *'Provisional Hypothesis of Pangenesis'* in his 1868 book *'Variation of Animal and Plants under Domestication'* (Darwin, 1868:357-404). His previous book, *'Origin of the Species'* (Darwin, 1859) received substantial attention in the scientific community (as well as in the wider public). *Origin of the Species* includes a chapter on *'Variation under Domestication'* (Darwin, 1859:7-43). The writings of Darwin on heredity were widely read and inspired numerous contemporary scientific works on heredity. Francis Galton (1822-1911) was the half-cousin of Charles Darwin, and a true pioneer for the theoretical work on heredity (Galton, 1875). Galton wrote his first contribution to this topic in 1865 (Galton, 1865). Galton challenged Darwin's heredity theory of Pangenesis (Galton, 1871a, 1871b), and also explicitly rejected the idea of acquired characters know as *'Lamarckism'* (Galton, 1876:346). Almost all of the most influential works on heredity in the late 1800s referenced the work by Darwin, and in particular his *'Hypothesis of Pangenesis'* (see for example Brooks, 1883; Nägeli, 1884, 1898; Vries, 1889; Weismann, 1893). Mendel's famous 1866 manuscript was not unknown or unavailable to contemporary scientists; it was referenced at least 10 times by contemporaries. However, only Karl Wilhelm von Nägeli (1817-1891) seemed to be aware of the aspects of Mendels work in relation to Heredity before 1900 - but Nägeli dismissed Mendel's conclusions as implausible (Kampourakis, 2010). It was William Bateson (1861-1926) who made Mendel's manuscript available for an English speaking audience (Mendel, 1866; Bateson, 1902). In 1910 Bateson became the first director of the John Innes Horticultural Institution (now John Innes Centre, JIC, http://www.jic.ac.uk, verified 5 Feb 2011).


Jean-Baptiste Lamarck (1744-1829)


Charles Darwin (1809-1882)


Francis Galton (1822-1911)

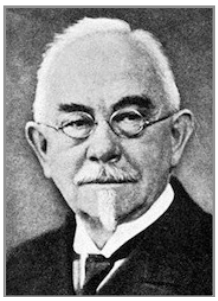

Friedrich Leopold August Weismann (1834-1914)


William Bateson (1861-1926)

One of the most important contributions to the understanding of Mendelian genetics came from a young student of eugenics, Ronald Aylmer Fisher (1890-1962). In 1916, he submitted, at the age of only 26, a groundbreaking manuscript to describe how traits measured on a continuous measurement scale is consistent with Mendelian principles, and in support of the concept known as saltations, to the Royal Society of London. Two very distinguished scientists, Reginald Crundal Punnett (1875-1967) and Karl Pearson (1857-1936) acted as referees, but found themselves unable to fully understand the manuscript. After a lengthy correspondence with referee Karl Pearson, Ronald Fisher submitted the manuscript to the Royal Society in Edinburgh instead where it was published in 1918 (Fisher, 1918). Fisher took up work as statistician to study crop variation at the Rothamsted Agricultural Research station located at Harpenden north of London, where he stayed for 24 years.


Luther Burbank (1849-1926)


Ivan V. Michurin (1855-1935)


Wilhelm L. Johannsen (1857-1927)


Nils Herman Nilsson-Ehle (1873-1949)


Ronald Aylmer Fischer (1890-1962)

Luther Burbank (1849-1926) achieved legendary results as a breeder of new cultivars from his nursery in Santa Rosa, California. He was praised as a wizard by the public for his ability to produce improved cultivars of fruit trees, vegetables and ornamentals. The working methods of Burbank was more those of improvisation. When George Harrison Shull (1874-1954) was commissioned to write a report on Burbank's work for the Cold Spring Harbor Laboratory, he found that Burbank kept very few records of his experiments and that his scientific approach was very unorthodox. Burbank was a firm believer in the theory of inheritance of acquired characteristics, and later came to be associated with the movement of Trofim Denisovich Lysenko (1898-1976). Another Russian (and *Lysenkoist*), Ivan Vladimirovich Michurin (1855-1935) can be seen as *'the counterpart of Burbank in the Soviet'*. The major contributions of Michurin were in pomology. For more information about Luther Burbank, see Crow (2001) and Stansfield (2006).

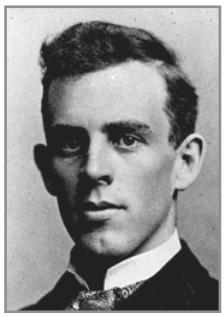
The Danish botanist Wilhelm Ludwig Johannsen (1857-1927) contributed works on plant physiology and genetics that was setting international standards in the beginning the 20th century. He is perhaps most famous for his works on the so-called *pure lines* of self-fertile common bean (*Phaseolus vulgaris* L.) (Kim, 1991). This discovery provided a relatively easy

way to create genetically identical plants for crop research. Johannsen introduced the terms *'genotype'* and *'phenotype'* (Johannsen, 1903) and the term *'gene'* that came to replace the *pangenesis* term coined by Darwin (Johannsen, 1905; Roll-Hansen, 2009). In Sweden, Nils Herman Nilsson-Ehle (1873-1949) became the front scientist of the prosperous movement of Mendelism, and he contributed to rapid advances in plant breeding based from his base at Svalöv in the south of Sweden (see for example Nilsson-Ehle, 1911, 1919). From Norway we could mention Werner Hosewinckel Christie (1877-1927) who was the first Norwegian to make a doctoral thesis in an agricultural topic (Christie, 1914). Christie showed that the diversity was richer in less cultivated forms using cereal, potato and pea plants as examples (Linnestad, 2001).

2.5 Early seed hunters and pioneer plant explorers

Frank Nicholas Meyer (1875-1918) is famous for collecting trips deep into Asia.

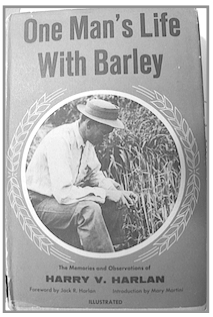
He was born Frans Nicholaas Meijer in Amsterdam and from the age of 14 worked under the direction of the legendary botanist Hugo de Vries at the Amsterdam botanical garden. At age 22 (in October 1901) his great wanderlust lead him to the US where he found work at the United States Department of Agriculture (USDA) under the direction of David Fairchild (1869-1954) at the USDA Arboretum. During 1902-1904 his first collecting trip lead him to Mexico, California and Cuba. In 1905 Fairchild sent him on an expedition (1905-1908) to collect plants in China. Meyer was dazzled by the diversity of plants he encountered in Asia and returned for another three collecting trips (1909-11, 1912-15, 1916-18). In 1912 China turned into a republic followed by a very unstable political period. These were dangerous times for Asian expeditions, and during his last expedition Meyer died on 28 May 1918 under somewhat mysterious circumstances sailing down the Yang River towards Shanghai. Meyer brought approximately 2500 plant introductions to the USA including soybeans, new types of grain, fruit, vegetables, bamboos, and a number of ornamental trees and scrubs (www.PlantExplorers.com, 1999-2010).



David G. Fairchild (1869-1954)

Frank N. Meyer (1875-1918)

Harry V. Harlan (1882-1944)



One Man's Life With Barley (H. Harlan, 1957).

Harry Vaughn Harlan (1882-1944) visited South America, Europe, Africa and Asia to collect and study barley (Harlan, 1957). He was not known as a writer, but the brief letters and notes from his expeditions provide valuable information of traditional agricultural practices beyond those of barley only. Harry Harlan was the father of Jack Rodney Harlan (1917-1998).

2.6 N.I. Vavilov and his institute in Saint Petersburg

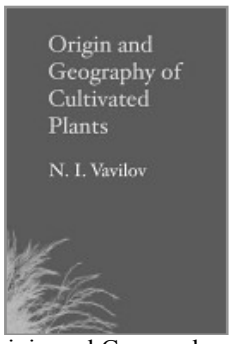
Nikolai Ivanovich Vavilov (1887-1943) grew up in Moscow as the son of a self-made and successful textile merchant. His brother Sergei became a famous physicist, his sister Alexandra a physician, and his second sister Lidia a microbiologist. After his studies in Agriculture, Vavilov worked for a short period during 1911-1912 at the Bureau of Applied Botany (predecessor to the Vavilov Institute). Robert E. Regel (1867-1920) was director of the Institute since 1905, and he became an important support for Vavilov in seeking a scientific career. In 1913 Vavilov traveled to Great Britain to study plant breeding and genetics under supervision of William Bateson. His first major collecting expedition to Asia led him to Iran and to the Palmir

Mountains. In 1917 Vavilov was appointed as a Professor in Agriculture at an institute in Saratov, and the same year also assistant head under Regel at the Department of Applied Botany (Bureau of Applied Botany, 1894-1916). When Regel died in 1920 Vavilov accepted to replace him as director.

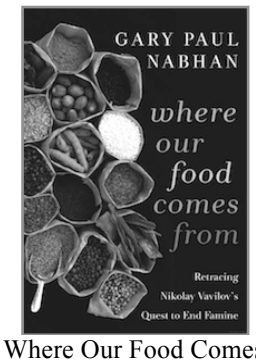
Under the leadership of Vavilov the Institute became one of the leading plant breeding and research institutes in the world, and Vavilov himself received numerous honorable awards and scientific Society membership appointments both in Russia and abroad. In his work Vavilov found great inspiration in the work by Candolle on the origin of cultivated plants and in that of Darwin for genetics: *"After reading De Candolle, Darwin and Gehn a dozen times, I feel quite deservedly that we have managed to get a bit ahead of De Candolle"* (Vavilov, 1923 *cf*. Esakov, 1980; *cf*. Cohen, 1982; *cf*. Loskutov, 1999:16). Vavilov produced an impressive number of scientific manuscripts and studies, including his work on the *'Law of Homologous Series in Variation'* (Vavilov, 1920, 1922) and *'Centers of origin of cultivated plants'* (Vavilov, 1917, 1924, 1926, 1940).

Nikolai I. Vavilov (1887-1943)

Vavilov and his Institute (Loskutov, 1999)



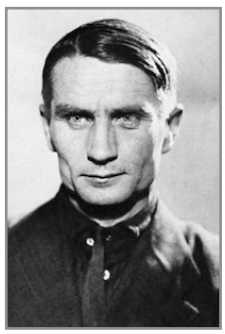
Origin and Geography of Cultivated Plants (Vavilov, 2009)



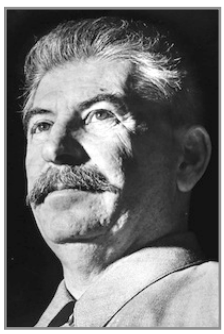
Where Our Food Comes From (Nabhan, 2009)

In 1926 Vavilov received one of the highest recognitions of the USSR: the Lenin Medal. Soviet agriculture became during the 1930s more and more infected by the influence of Trofim Lysenko. Lysenko was an unknown agronomist before he made claims that his vernalization method for wheat three to four times increased agricultural yields (Lysenko, 1951). Lysenko further claimed a Lamarcksist basis in that the dramatically improved yield his method could produce was passed on the next generation in only one single *'breeding'* step. The theory seemed to find ground in the Marxist Soviet leadership, and soon any scientific objection to Lysenko's results, not successfully replicated by anybody, was censored. Lysenko gradually replaced Vavilov in all influential positions in USSR agriculture, and in 1940 Vavilov was secretly arrested during a collecting trip in the Caucasus - and disappeared. Not even his family knew his whereabouts. On 26 January 1943 Vavilov died of malnutrition in a prison cell in Saratov. It would take another 20 years before any official details of these events were released.

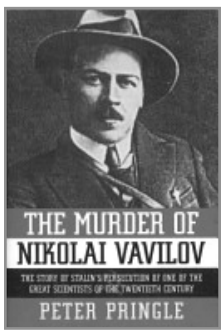
Soviet Union (1922-1991)



Trofim Denisovich Lysenko (1898-1976)



Joseph Vissarionovich Stalin (1878-1953)



The Murder of Nikolai Vavilov (Pringle, 2009)

The *'Bureau of Applied Botany'* (1894-1916) was, in 1916, renamed the *'Department of Applied Botany and Plant Breeding'* (1916-1924); and again in 1924 to the *'All-Union Institute of Applied Botany and New Crops'* (1924-1930). In 1924 the Institute was reorganized as the *'All-Union Institute of Plant Industry'* (1930-1992). The current name, *'N. I. Vavilov Research Institute of Plant Industry'* was given in 1992 to honor of the achievements of Nikolai Vavilov. For further reading about N.I. Vavilov and his institute, the book by Igor Loskutov from 1999 provides an excellent overview. A few of his selected works have been translated to English including *'Five Continents'* (Vavilov, 1997) and *'Origin and Geography of Cultivated Plants'* (Dorofeev, 1992). Of more recent production, *'The Murder of Nikolai Vavilov'* (Pringle, 2008) reads like a real crime thriller, and *'Where our food comes from, retracing Nikolay Vavilov's Quest to End Famine'* (Nabhan, 2009) provides an accessible story of the Vavilov's work in the context of our present time.

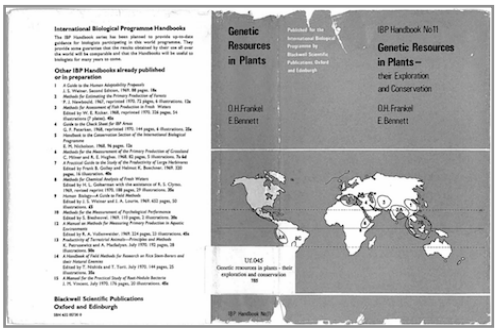
My first visit to the Vavilov Institute March 2002, photo by Dag Endresen

My second visit to the Vavilov Institute May 2003, photo by Dag Endresen

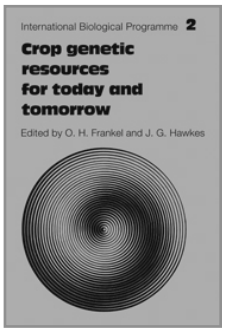
My third and latest visit to the Vavilov Institute April 2010, photo by Dag Endresen

2.7 Establishment of Vavilovian genebanks in the 1960s

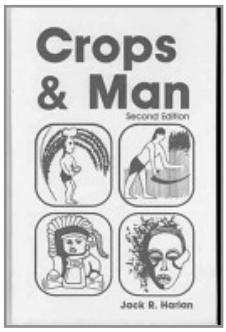
Before Vavilov, landraces and other plant genetic resources were collected and introduced as is, or included in plant breeding programs. Germplasm material was maintained by botanical gardens, but few efforts were made to regenerate and conserve the material. The collection and introduction of germplasm was for immediate use, and not for long-term conservation (Thompson, 1970). Vavilov collected germplasm material to build a large long-term reference genebank collection, called the *'world collection'* by his colleagues. Evaluation and charactization of the material in this large collection provided the basis for research experiments and for the development of new and improved crop cultivars (Frankel, 1988). Harry Harlan and Mary Martini wrote about the loss of landraces already in the 1930s (Harlan and Martini, 1936). But it was the writings of Otto Frankel on genetic erosion since the 1950s that sparkled the establishment of many large genebank collections for long-term conservation following the *Vavilovian* principles (Frankel, 1950). For further reading on the *"discovery"* of the Vavilov concept of genebanks for long-term conservation of plant genetic resources, I recommend the following books, available for a wider audience. In 1970 *"Genetic Resources in Plants - their Exploration and Conservation"* was published as a compilation of contributions to the 1967 conference organized by the Food and Agriculture Organization of the United Nations (FAO) (Frankel and Bennett, 1970). Otto Herzberg Frankel (1900-1998) and Erna Bennett (1925-) were the editors for this book, and it was often referred to as the *"genebank bible"* in the following years. As a product of the International Biological Program (IBP, 1964-1974) the *"Crop genetic resources for today and tomorrow"* (Frankel and Hawkes, 1975) became another important book for the new genebanks that were established in the 1960s and 1970s. *"Crops and Man"* (Harlan, 1975, 1992) is a study on the origin of the food crops. *"Genes, Crops and the Environment"* (Holden et al., 1993) was a somewhat more recent book, and more accessible to the general public.



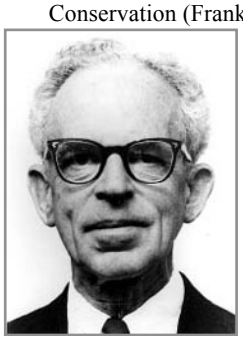
Genetic Resources in Plants - their Exploration and Conservation (Frankel and Bennett, 1970)



Crop genetic resources for today and tomorrow (Frankel and Hawkes, 1975)



Crops & Man (J.R. Harlan, 1975/1992)



Otto H. Frankel (1900-1998)

Erna Bennett (1925-)

John G. Hawkes (1915-2007)

Jack R. Harlan (1917-1998)

2.8 Breeders' rights, UPOV, TRIPS

The fist steps towards modern plant breeding were made in the middle of the 19th century. Public agricultural research stations performed the initial plant breeding. After the re-discovery of the work by Mendel and his principles of heredity, the progress in increasing agricultural yield took off, and private breeding companies like Pioneer Hi-Bred started to replace public funded plant breeding (Kloppenburg and Kleinman, 1987). The private seed industry started to demand *Breeders' Rights* for ownership of the new cultivars they develop (Weibull, 1997). The Plant Variety Protection Act of 1970 (PVPA) marked a shift as it grants (to plant breeders in the USA) intellectual property rights for a period of 25 years of new cultivars that are novel, stable and uniform (Chen, 2006). Some voices started to raise concerns for the protection of *Farmers' Rights* to the genetic resources developed during centuries or even millennia of farming (Fowler and Mooney, 1990).

2.8.1 Union for the Protection of New Varieties of Plants (UPOV)

In 1961 the International Union for the Protection of New Varieties of Plants (UPOV) was established at an international conference held in Paris (UPOV entered into force on 10 August 1968). The UPOV regulations established an international framework for the protection of plant breeders' rights to the cultivars they develop (UPOV, 1991). The primitive cultivars, landraces and other plant genetic resources represented by the crop wild relatives were at the time still considered as common human heritage.

2.8.2 Trade Related Aspects of Intellectual Rights Agreement (TRIPS)

The Trade Related Aspects of Intellectual Rights Agreement (TRIPS) was agreed in 1994 during the final rounds of the international meetings to pass from the previous General Agreement on Tariffs and Trade (GATT, 1949-1993) to the new World Trade Organization (WTO, 1994). The TRIPS came into force in 1995 and is a requirement for membership in WTO. The TRIPS provides novel legislation enabling the patenting of biological life forms, including new cultivars developed by plant breeders. This can be seen as an enhancement of the *Breeders' Rights* and caused understandably a renewed focus on the protection of *Farmers' Rights* (see for example Ramanna, 2006). In 2001 the Indian Government passed a new

legislation for *The Protection of Plant Varieties and Farmers' Rights Act* (Brahmi et al., 2004; Koo et al., 2004).

2.9 The agriculture sector at FAO versus the environment sector at UNEP

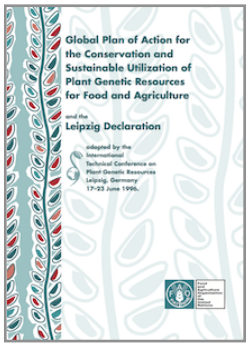
The *International Institute of Agriculture* (1908-1945) located in Rome was the predecessor to FAO (International Convention, 1910, 1930). After the establishment of the United Nations during World War II, the functions of the International Agricultural Institute were in 1946 transferred to the newly started Food and Agriculture Organization of the United Nations (FAO) (United Nations, 1948). The International Undertaking (IU), Global Plan of Action (GPA) and the International Treaty (ITPGRFA) are conventions developed at FAO (agriculture sector). A parallel process for similar international legal instruments was running in the environment sector, at the United Nations Environment Programme (UNEP). The Convention of Biological Diversity (CBD) and the Access and Benefit Sharing (ABS) framework was developed at UNEP.

2.10 Farmers' rights and common heritage of humankind (IU, 1983)

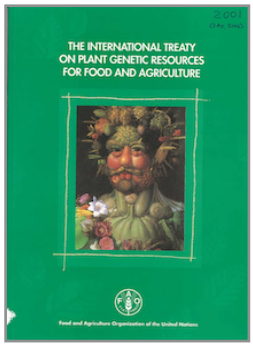
In November 1983 the 22nd FAO conference agreed on a resolution (8/83) to establish the International Undertaking on Plant Genetic Resources for Food and Agriculture (IU). The IU was a non-binding agreement and stated that plant genetic resources [as conserved and made available by genebanks] were the common heritage of humankind. Some countries, including Denmark, Finland, Norway and Sweden, expressed concern that the IU explicitly included regulations for: *"special genetic stocks (including elite and current breeders' lines and mutants)"* (FAO, 1983) and officially declined to support the IU (Kloppenburg and Kleinman, 1987). The IU recognized *Farmers' Rights* as complementary to *Breeders' Rights*, and further that both types of germplasm material were to be considered as a common heritage of mankind. The IU can be seen as limiting the rights of the breeders to claim ownership over the commercial cultivars that they develop and release. The IU was officially supported by a total of 113 countries (FAO, 2001).

FAO (1910, 1948)

International Undertaking (1983)

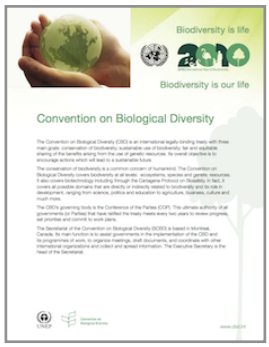


Global Plan of Action (1996)

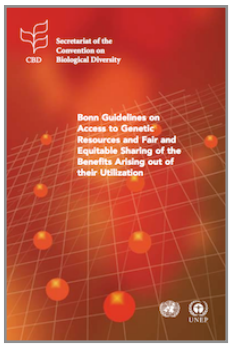


ITPGRFA (2001/2004)

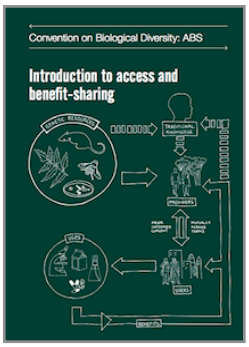
UNEP (1972)



CBD Rio (1992/1993)



CBD ABS, Bonn (2002)



CBD ABS Nagoya (2010)

2.11 Biodiversity as national sovereign property (CBD, 1992)

In 1988 the United Nations Environment Programme (UNEP) initiated the work on a legal instrument for conservation and sustainable use of biological diversity. After a series of 6 preparatory international meetings during 1991 and 1992 in Spain, Kenya, and Switzerland, the Convention of Biological Diversity (CBD) was opened for signature at the Earth Summit 5 June 1992, Rio de Janeiro, Brazil. This new legally binding treaty came into force on 29 December 1993 with the ratification by 168 countries. The CBD prescribe the concept of *'national sovereign rights over their own biological resources'* (United Nations, 1993:preamble) and the concept of *'sharing in a fair and equitable way the results of research and development and the benefits arising from the commercial and other utilization of genetic resources'* (United Nations, 1993:Article 15). The principles for access and benefit sharing (ABS) were developed further following the establishment of a working group subsidiary to the convention in 2000 (at the 5th meeting) and the adoption of the Bonn guidelines in 2002 (United Nations, 2002). The underlying principle of the CBD is that countries have the right and obligation to decide the fate of their biodiversity. The observed loss of biological diversity has caused an alarm of potential devastating effects for human welfare (Wilson, 1992; Diamond, 2005). By providing national sovereign rights and ownership to biodiversity and genetic resources the hope is that biological diversity will be perceived as having a more real value, and that this will stimulate to the sustainable use and a more responsible attitude to the conservation of these resources.

2.12 Human mediated exchange of cultivated plants

The CBD required a revision of the International Undertaking (IU) for the cultivated plants. The definition of national sovereign rights to biodiversity does not implement easily for the cultivated biodiversity. Some of the most typical plants exploited in national cuisines often originate from an entirely different continent. The potato (*Solanum tuberosum* L.) so essential in northern Europe was developed in the Andes, primarily in present Peru thousands of years ago. The potato only arrived in Europe after Columbus. The Italian kitchen is hard to imagine without thinking of the tomato (*Solanum lycopersicum* L.) that also arrived in Europe after Columbus. Maize (*Zea mays* L.) originally from present day Mexico is grown all over the world including Africa were maize and cassava (*Manihot esculenta* Crantz) from South America, have come a long way towards dominating the African staple crops. Wheat (*Triticum* ssp.) is the most important staple crop worldwide, including USA, Canada, Argentina and Brazil. Even if wheat spread with human settlements and pre-historic interaction between cultures, this crop is not assumed to have arrived in America before Columbus, and into large-scale cultivation much later (Simmonds, 1976; Smartt and Simmons, 1995; Diamond, 1997; Tannahill, 2002). Wheat provides a good example of a crop with a collaborative cultivation history. What country can claim sovereign property rights to cultivated plants? Does the ownership belong to the present countries located where the crop was first domesticated, also called *'country of origin'*? Does the ownership belong to the countries where the diversity of the crop itself or perhaps its wild relatives is larger, also called *'center of diversity'*? Or does the ownership simply reside with the country that collected the seeds from these crops before the CBD went into force? Can anybody claim ownership of something developed during pre-historic times? The copyrights for intellectual property rights most often expire after 50 or 70 years (Berne Convention, 1886).

2.13 Global Plan of Action (GPA, 1996)

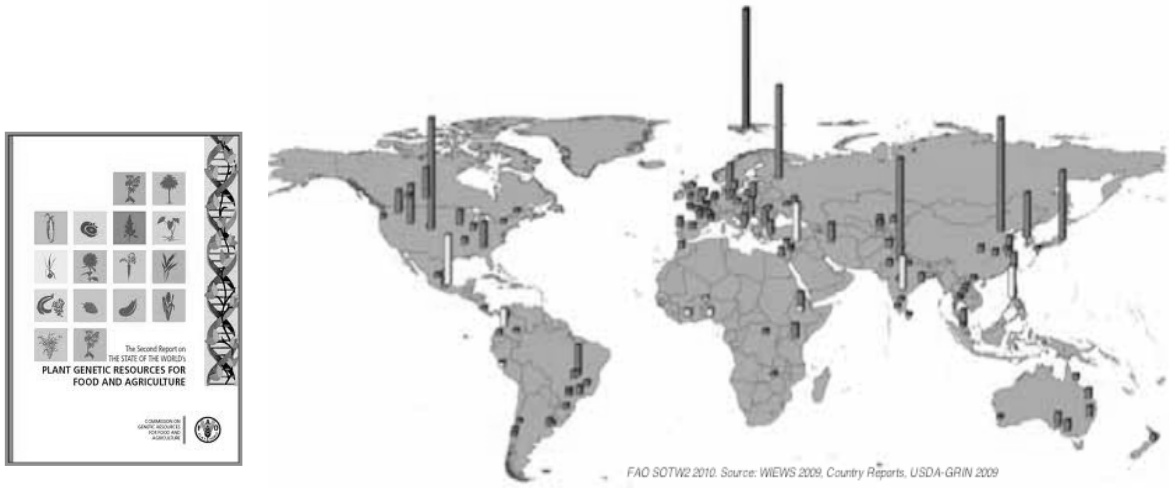
The *Global Plan of Action for the Conservation and Sustainable Utilization of Plant Genetic Resources for Food and Agriculture* (GPA) can be seen the first response to the CBD from the genetic resources sector. The GPA (FAO, 1996b) defines a set of recommendations and activities following the first *'Report on the State of the World's Plant Genetic Resources for Food and Agriculture'* (FAO, 1996a, 1997). The GPA was formally adopted as a government-level commitment through the *'Leipzig Declaration'* (FAO, 1996b).

2.14 International Treaty (ITPGRFA)

Following the GPA a natural next step was to develop a legally binding convention for plant genetic resources, similar to the CBD. Based on the International Undertaking (IU) and the Global Plan of Action (GPA) the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) was ready in 2001, and adopted at the 31st session of the Conference of the Food and Agriculture Organization of the United nations (31st FAO conference) on 3 November 2001 (FAO, 2002). The treaty came into force on the 90th day after the official ratification by 40 countries, on 29 June 2004 (FAO, 2009). The IU, GPA, and ITPGRFA can be seen as a protection of farmers' rights. The ITPGRFA implements a mechanism for the access and benefit-sharing framework introduced by the CBD. A major setback for the *'plant treaty'* was the absence of important crops like soya, sugar cane, oil palm and groundnut. The negotiations leading to the so-called *'Annex 1'*, listing the included crops were rough. Different countries made strong arguments for a more restricted or a more inclusive list of crops (Fowler et al., 2003).

2.15 Second Report on the State of the World's PGR (SOTW2)

The Food and Agriculture Organization of the United Nations (FAO) published in the fall of 2010 The *Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture* (SOTW2; FAO, 2010). The SOTW2 provides a comprehensive overview of the status for plant genetic resources. The report is based on the country reports describing the status for conservation and use of plant genetic resources in each country. FAO received in total 109 country reports for the preparation of the SOTW2.



The State of the World's Plant Genetic Resources for Food and Agriculture (FAO, 2010)

2.16 Access and Benefit Sharing (ABS, CBD Nagoya 2010)

The *Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity* (http://www.cbd.int/abs/, verified 25 Jan 2011) is the latest development from the CBD of particular relevance for the plant genetic resources sector. The *Nagoya Protocol* (ABS) defines an *international regime* for: *'fair and equitable sharing of the benefits arising out of the utilization of genetic resources'* (United Nations, 2010) and is based on the Bonn Guidelines (United Nations, 2002). The purpose of the ABS is to ensure that biodiversity-rich countries receive a fair monetary compensation for the use of genetic resources from their country. The ultimate purpose is poverty reduction and a mechanism to support and encourage sustainable use of genetic resources.

CHAPTER 3

3. The use of plant genetic resources

The establishment and maintenance of germplasm genebank collections to preserve the plant genetic diversity of crop plants provides an important function (Plucknett, 1987), but also challenges as it occurs financial costs and commitments for maintaining the genebank collection (Koo, 2004). The establishment of genebank collections implies a long-term commitment to maintain and develop the collection (Yndgaard and Kjellqvist, 1984). These commitments are not compatible with political instability and changing political priorities as demonstrated by the current budget cuts experienced by most genebanks in present times (Löfgren, 2008). Perhaps it is too easy for countries to start ever increasing numbers of new national genebank collections without a second thought to the long-term commitments in regards to financial and political support these collections require (Fowler, 2010). The genetic resources preserved and made available from the genebanks is a resource of high economic value and their exploitation generates today a much higher economic payback than the operation cost for these facilities (Rubenstein et al., 2005; Johnson, 2008; Smale et al., 2001).

3.1 Why do we need genebanks?

In situ and on-farm conservation is often considered the optimal strategy for maintaining valuable genetic diversity (Maxted et al., 1997, 2008; Iriondo et al., 2008; Brush, 2000; Jarvis et al., 2007). *In situ* and on-farm the genetic resources are in their natural environment and will be able to adapt and respond to changing conditions. But in modern times we have witnessed a gradual loss of genetic diversity with the (still ongoing) replacement of landraces and traditional cultivars by modern, high-yielding cultivars. The modern cultivars are more uniform than the previous landraces (Zeven, 1998). Genetic uniformity is even a requirement for the registration of new cultivars (UPOV, 1991). The loss of suitable habitat for the crop wild relatives has caused concerns regarding their survival *in situ* (Iriondo et al., 2008). *Ex situ* backup may thus be required to rescue these genetic resources. Moreover, the genetic diversity in crop wild relatives are much more readily available for exploitation when plant breeders and other users can send a seed request to a genebank, rather than engage in collecting expeditions to get access to their useful genetic traits (Engels et al., 2008: 175-176). Some of the most important areas of genetic diversity, including Afghanistan and Iraq, are currently dangerous for a collecting expedition to enter, because of ongoing wars and political unrest. The genebank collections also includes obsolete cultivars and provide access to the genetic diversity represented by these germplasm materials after they leave the commercial seed trade.

3.2 Maintaining crop resistance to pests and diseases

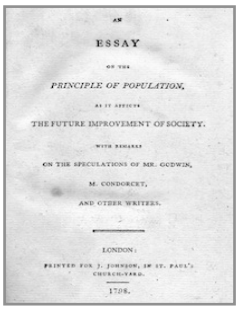
When modern cultivars were developed from the old landraces they passed through a so-called *"genetic bottleneck"* in respect to allelic diversity (Tanksley and McCouch, 1997; Doebley et al., 2006). Modern cultivars are optimized for high yield under a defined environment. Plant breeding selected for a limited range of genetic diversity to achieve a higher crop performance under specific environment conditions, including contemporary pathogens and crop pests. However the crop pathogens evolve and the genetic protection that was developed in the past is gradually broken down. With a limited genetic base in their working collections, plant breeders need access to novel genetic diversity to maintain the *'arms-race'* against crop pests and diseases. Influx of genetic diversity is needed to maintain the crop performance under the challenges of a changing world. Modern mechanized agriculture is characterized by large land areas planted with only one single crop (monoculture). Modern plant breeding generally target

uniform plant height and synchronous ripening to ease mechanized harvesting. This breeding target causes a tendency of genetic uniformity. Monocultures and genetic uniformity can lead to increased risk of crop epidemics with devastating impacts. A global pattern of increased human travel and interaction can assist the new crop pathogens and pests to spread. These are all important elements for a strategy to safeguard future agricultural performances (Qualset and Shands, 2005).

In the early 1970s the USA experienced a major crop epidemic when a fungal pathogen called southern corn blight (*Bipolaris maydis*) destroyed nearly $1 billion worth of the US corn harvest. More than 50% loss was reported in the worst affected states. Vast uniform stands of corn with almost identical maternal genotypes could be one of the reasons for the large impact of this crop epidemic (Ullstrup, 1972; Balint-Kurti and Johal, 2009). Similar examples of crop epidemics send a warning that this is a threat we need to show attention to. To limit such crop epidemics in the future, it is paramount to keep the valuable genetic diversity present in landraces and wild relatives available for crop improvement programs. It is unlikely that selection in the working collections of the breeder in a short time can develop novel adaptation in the crop to these types of problems (McCouch, 2004). The genetic diversity already available in the landraces and wild relatives is expected to be the most readily available genetic resource, and also the most efficient genetic resource we have available to combat crop pathogens (Feuillet et al., 2007). Even if the landraces and wild relatives bring problems due to their more primitive genetic backgrounds (causing lower yields), the genetic diversity they hold is unique and the result from adaptations to their environment during a long time period. It is this genetic diversity that provides the raw material for active plant breeding and crop improvement research (Acquaah, 2007).

3.3 Human population growth will require increased food production

The human population is still growing fast (United Nations, 2004, 2007). Thomas Rupert Malthus (1766-1834) was one of the first to discuss the implications of extreme population growth (Malthus, 1798) that we currently witness. Novel solutions in agriculture are required to meet the demand from the increasing number of people wanting to eat. The genetic resources maintained by germplasm genebanks provide the raw materials for crop improvement and will thus play a crucial role in shaping the future (Motley et al., 2006). Mass-starvation of unseen proportions is not a completely unthinkable outcome of the present trend, even if biologist Paul R. Ehrlich predated the collapse in food supply in his 1971 book: *"The battle to feed all of humanity is over. In the 1970s and 1980s hundreds of millions of people will starve to death in spite of any crash programs embarked upon now. At this late date nothing can prevent a substantial increase in the world death rate"* (Ehrlich, 1971: introduction, page xi). Problems of food supply caused by population growth may possibly be available from our pre-historic history as well. Cohen (1977) proposes that population growth might be one of the reasons for the origin of agriculture.


AN
ESSAY
ON THE
PRINCIPLE OF POPULATION,
AS IT AFFECTS
THE FUTURE IMPROVEMENT OF SOCIETY.
WITH REMARKS
ON THE SPECULATIONS OF MR. GODWIN,
M. CONDORCET,
AND OTHER WRITERS.
LONDON:
PRINTED FOR J. JOHNSON, IN ST. PAUL'S CHURCH-YARD.
1798.

An essay on the Principle of Population (Malthus, 1798)

Thomas Rupert Malthus (1766-1834)

Photo by Dag Endresen, October 2010

3.4 The green revolution (Norman Borlaug)

Norman Ernest Borlaug (1914-2009) started his work at the newly established *'Cooperative Wheat Research and Production Program'* in Mexico in 1944. This *Program* was later reorganized as CIMMYT (*Centro Internacional de Mejoramiento de Maíz y Trigo*) in 1963 - under the leadership of Borlaug. The new institute in Mexico implemented a method they called *'shuttle breeding'*. Wheat lines were *"shuttled"* between the Yaqui valley in the north of Mexico and the Toluca valley in the south to achieve two growth seasons each year and thus cutting the breeding time in half. The wheat breeding station in the Mexican Yaqui valley was also integrated with agricultural research stations in the USA and Canada to provide a multinational crop research program to combat the wheat stem rust epidemics running up and down the west coast of North America each year in the early 1950s.

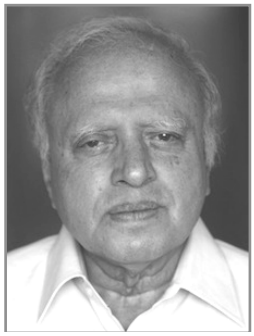
3.4.1 Multilines and semi-dwarf wheat

In 1953 Borlaug mixed multiple *pureline* wheats with different alleles for the most important stem rust resistance genes together as a so-called *"multiline"* mixture. This approach combined the efficiency of *pureline* breeding methods together with resistance to multiple races of stem rust (Borlaug, 1954, 2007). The same year in 1953, Borlaug acquired a *semi-dwarf* cultivar of wheat with Japanese origin called *'Norin 10'*. The semi-dwarf character means that the plants grow to almost half the plant height as standard wheats. This has important implications for the yield. Standard wheats selected for higher yields have a tendency to lodging. Lodging is a term to describe the tendency for tall plants to break in the wind under the weight of the big heads. With the dwarfing gene, breeding efforts can proceed to higher yields, less inedible plant material relative to the harvested grains, with substantially fewer problems of lodging. With lodging tackled, fertilizer levels could be increased and giving even higher yields. Borlaug started breeding efforts to combine the disease resistant *multilines* with the *dwarfing* character. The results of these breeding efforts resulted in the start of a legendary increase in yield known as the *'green revolution'*. Borlaug is said to have saved more then one billion people from starvation, and in 1970 he received the Nobel Peace prize for this achievement (Nobel Foundation, 1970).

3.4.2 The green revolution in India

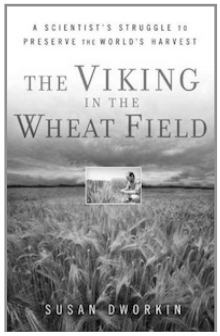
In 1962, Monkombu Sambasivan Swaminathan (Tamil: மான்கொம்பு சாம்பசிவன் சுவாமிநாதன்) made the initiative for Borlaug to visit India to explore together the possibilities for implementation of Borlaug's' results in India. It was the enormous progress of high-yielding *'Mexican wheats'* in India and Pakistan that gave rise to the well-known term: *'green revolution'*.

Norman Ernest Borlaug (1914-2009)



M.S. Swaminathan (1925-)

Bent Skovmand (1945-2007) Photo Dag Endresen 2005



The Viking in the Wheat Field (Dworkin, 2009)

3.5 Sir Bent Skovmand

Bent Skovmand (1945-2007) was responsible for starting my PhD. He was one of the talented young scientists that Borlaug called upon to join the team in Mexico (Dworkin, 2009; Zeyen and

Groth, 2009). Skovmand was an expert on stem rust by training (Skovmand, 1973, 1976). One of the most important contributions by Bent for genetic resources was the development of the CIMMYT pre-breeding lines where he transferred useful traits from the old landraces to a more modern genetic background. Author Susan Dworkin described these contributions in her latest book: *'The Viking in the Wheat Field'* (Skovmand et al., 2001; Dworkin, 2009). See appendix 2 for further details.

3.6 Did the green revolution cause genetic erosion?

The Green Revolution led to an impressive increase in food production, but could this increase of agricultural yield have less attractive side effects? The Green Revolution is sometimes set in connection with environmental problems and issues of sustainability. The recent developments in industrial agriculture created large fields of homogenous, genetically uniform crops. Weeds and pests are efficiently removed with chemical pesticides and weed killers. We have created large agricultural areas with room for only two species: humans and the species we cultivate (*'duocultures'*). Recent studies report that we apply fertilizers and use mechanized farming practices with an estimated higher total energy input than the equivalent energy output for many of the important crops (Hatirli et al., 2005; Heidari and Omid, 2011). The replacement of landraces by improved cultivars developed by plant breeders was observed long before the onset of the Green Revolution (Proskowetz, 1890; Baur, 1914; Harlan and Martini, 1936; Frankel, 1950). Unsustainable land management and cultivated monocultures are not rare events in previous societies and civilizations (Diamond, 2005). Melinda Smale (1997) argued that the Green Revolution was not responsible for genetic erosion and genetic vulnerability in wheat. Improved wheat cultivars showed reduced susceptibility to rust diseases, and the understanding of the genetic basis of resistance steadily increase. Borlaug (2007) argued that the total land area used for food production is only marginally larger than it was before the 1960s, while the harvested crops increased almost three times. Borland reminded us that this enormous increased agricultural yield could be seen as sparing large areas of natural habitat from transformation to cultivated lands. The improved wheat cultivars are more responsive to increased levels of fertilizers applied, and substantially more effective in transforming fertilizer inputs to improved yields. Of equal importance is the improved agricultural production under lower irrigation. The increased efficiency in respect to water is the result of both new cultivars and new farming methods. *"[H]unger has dropped dramatically since the 1960s - from 40% of world population to about 17% - because of rapid population growth there are still 850 million hungry people (...) Too many people live in hunger and sickness. As Lord John Boyd Orr, the first director general of FAO and a Nobel Peace Laureate said in 1950 in his Nobel acceptance speech, 'peace cannot be built on empty stomachs' to which I add 'and human misery'"* (Borlaug, 2007:297).

3.6.1 Molecular methods to sustain increased food production

Molecular methods to develop genetically modified crops cause a similar public concern in respect to environmental impacts. For reasons I never understood, other concerns are more focused on the potential harm to human health posed by these GMOs (genetically modified organisms). Molecular methods have demonstrated great potential for crop improvements with respect to a wide diversity of traits including yield (Tanksley and McCouch, 1997; Davies, 2003; McCouch, 2004). Molecular plant breeding methods and GMO crops could contribute to avoid large-scale hunger to escalate. The latest estimates on human population growth indicate a balance at approximately 9.2 billion people around 2075 (United Nations, 2004, 2007). While we wait for the peak in 2075, there will still be added 34% more people to the present world population of 6.87 billon (http://math.berkeley.edu/~galen/popclk.html, visited 28 Jan 2011). Increased agricultural yields are still required to avoid large-scale hunger and destruction of

natural habitats. If the GMO crops show the most promising potential to achieve increased yields, we need very strong arguments to stop the implementation of these new technologies.

3.7 Global warming and climate change will add new challenges

The Intergovernmental Panel on Climate Change (IPCC) has issued a warning that the global climate is warming at an alarming rate (IPCC, 2007). Past times of rapid climate change has been proposed as one of many hypotheses for the origins of agriculture (Childe, 1952). Using the MaxEnt species distribution model Jarvis et al. (2008) estimated the effect of climate change for the wild relatives of peanut (*Arachis*), potato (*Solanum*) and cowpea (*Vigna*). The availability of suitable habitat in a future climate scenario for year 2055 (Govindasamy et al., 2003) compared to the present climate (Hijmans et al., 2005; http://worldclim.org) was reduced with more than 50% for most of the species they modeled. Particularly alarming was the predicted extinction of 16-22% of these species. With the wild relatives of the cultivated plants predicted to suffer dramatic loss of diversity, the situation for the plants we cultivate and rely on for food subsistence is not much better. A recent report from the FAO Commission on Genetic Resources for Food and Agriculture (CGRFA) describes how the current period of rapid climate change is expected to cause dramatic effects on biodiversity including agrobiodiversity and agricultural food production (Fujisaka et al., 2009, 2010). The demand for increased food production is already at a very high level because of extreme population growth (United Nations, 2004, 2007). If the challenges from climate change lead to a lower level of food production the outcomes could be severe (Cribb, 2010).

3.8 Genetic Erosion

The *Vilmorin-Andrieux et Cie* [Vilmorin Breeding Institute] was established in 1727 and can be seen as the very first beginning of plant breeding (Anonymous, 1930:224; Acquaah, 2007:7). The disappearance of the old landraces was noticed shortly after the further expansion of systematic plant breeding in Europe in the 19th century. The Czechoslovakian plant breeder, Emanuel von Proskowetz, called for the useful genetic diversity represented by the landraces in Europe to be collected, described, documented and preserved (Proskowetz, 1890 *cf.* Lehmann, 1981). Baur (1914 *cf.* Lehmann, 1981) added to this recommendation to collect and preserve landraces from outside Europe, and their wild relatives. Later Harry Harlan and Mary Martini repeated the warning that the loss of available genetic diversity caused by the replacement of landraces with the more uniform modern cultivars required increased attention: *"In the great laboratory of Asia, Europe, and Africa, unguided barley breeding has been going on for thousands of years. Types without number have arisen over an enormous area. The better ones have survived. Many of the surviving types are old […] the progenies of these fields with all their surviving variations constitute the world's priceless reservoir of germplasm. It has waited through long centuries. Unfortunately, from the breeder's standpoint, it is now being imperiled. When new barleys replace those grown by farmers of Ethiopia and Tibet, the world will have lost something irreplaceable. When that day comes our collections, constituting as they do but a small fraction of the world's barley, will assume an importance now hard to visualize"* (Harlan and Martini, 1936:317; *cf.* Harlan, J.R., 1975:618-619; Harlan, J.R., 1995:244; and Kloppenburg, J.R, 1988:162). In 1950 Frankel repeated the warning on the loss of landraces (Frankel, 1950), but his warning did not invoke any major activities in the 'Western world' with respect to the conservation of germplasm in genebanks.

The 1960s was a decade of increasing international focus on the need to conserve plant genetic resources (Frankel and Bennett, 1970). This lead in 1974 to the establishment of the International Board for Plant Genetic Resources (IBPGR) with a secretariat located in Rome hosted by FAO (IBPGR, 1974; Harlan, 1975; Lehmann, 1981). IBPGR was established to initiate and coordinate international activities on the collection and documentation of plant

genetic resources with a special focus on the old landraces (Lawrence, 1984). More recently Wouw et al. (2009) conducted an analysis on the genetic erosion in crops and found that the replacement of landraces with modern cultivars most likely has caused a reduction in crop diversity, but also that the genetic erosion did not continue with further development of the modern cultivars. They further commented that the replacement of landraces with modern cultivars seems to have been more characteristic for North America and Western Europe than for many other parts of the world. A second literature study including 44 published papers conducted by the same authors focused on the genetic diversity trend in modern cultivars during the previous century (Wouw et al., 2010). They found again no clear evidence for the continued gradual loss of genetic diversity in modern cultivars during this period. Some years earlier A. Kolodinska Brantestam (2005; Kolodinska Brantestam et al., 2004) came to similar conclusions with her PhD thesis from 2005 based on molecular marker studies of genetic diversity in Nordic barley also during the 20th century. However Hammer and Diederichsen (2009) commented that the results from such recent genetic diversity studies using molecular approaches might draw ambiguous conclusions. This is because such studies are often based on non-functional genetic diversity, and older landraces are under-represented in such studies.

3.9 Utilization of genetic diversity from landraces and wild relatives

The initial productivity gain in agriculture came from increasing the cultivated land area. The second productivity gain came from improved land productivity (yields). After more than a century of very successful plant breeding achievements it will now be more important than ever to explore new sources for the trait enhancing alleles (Ruttan, 2002). The most common classification of the cultivated plants separates modern cultivars, obsolete cultivars, landraces, wild relatives, genetic stock, and breeding lines (Frankel, 1977). The modern and obsolete cultivars have been exhaustively explored as sources of valuable traits for plant breeding. Meeting the future demands of crop improvements will require increased exploration of novel genetic diversity from landraces and crop wild relatives (Skovmand et al., 2001). Hoisington et al. (1999) point to the challenges of accurate comparative evaluation of landraces due to intra-accession variability. When a useful trait is identified, the subsequent gene transfer to breeding lines and the modern cultivars creates new challenges. Wild relatives of the crop plants are even more genetically remote from modern cultivars and pose major challenges to utilization because of traits like seed shattering (Harlan, 1992). Modern molecular methods have improved the techniques for gene transfer (Tanksley and McCouch, 1997; Gepts, 2006), but the challenges of identifying valuable traits and the corresponding alleles remains an important issue to tackle. Allele mining has been proposed as a method to utilize the biotechnology advances of molecular biology (Tanksley and McCouch, 1997; Prada, 2009; Hamblin et al., 2010; Bhullar et al., 2010a). Bhullar et al. (2009, 2010b) described how to apply trait mining with FIGS as a complement to allele mining to find new sources for pest resistance for breeding (wheat powdery mildew). Trait mining with FIGS has been proposed as a strategy to improve the target identification of useful traits in landraces and crop wild relatives (Mackay and Street, 2004). With the development of new tools and methods to unlock the genetic diversity of landraces and crop wild relatives (Bhullar et al., 2009; PAPER II, PAPER III), the genetic resources held by genebanks will provide an attractive gene pool for novel alleles in future crop improvement. *"Without variability, it is not possible to conduct a plant breeding program! Germplasm is hence the critical first step in initiating a breeding program"* (Acquaah, 2007:73).

3.10 Pre-breeding to make plant genetic resources more available

If allele mining and trait mining approaches can assist with identification of useful genetic diversity, there still remains one more significant obstacle for the plant breeder wanting to bring traits from landraces or wild relatives of the cultivated plants into new cultivars. The genetic background of modern cultivars has undergone significant *'cleaning'* to arrive at more

homogenous and so-called *'true-breeding'* crops (e.g. *pure lines*, Johannsen, 1903; Acquaah, 2007:281-350). These modern cultivars and breeding lines are often called *'adapted'* materials and germplasm such as landraces *'unadapted'*. The plant breeders need the genetic diversity only from a few selected trait properties such as resistance to pathogens or the tolerance to new climate types. But classical plant breeding (crossing) will transfer much more of the genetic background from the landraces (or wild relatives) than the plant breeder desires. A series of time consuming so-called *'backcrossing'* efforts are required to restore the desired properties of the breeders line back to the qualities and yield performances expected from modern cultivars (Acquaah, 2007:101-102). Intra-accession variability with wide differences in many different traits for landraces and wild relatives makes both the comparative evaluation difficult and the subsequent gene transfer challenging (Hoisington et al., 1999). This process of bringing exotic traits into a more modern genetic background is often referred to as *'pre-breeding'* (see for example Bothmer and Linde-Laursen, 1989). Experiences from recent public-private partnerships on pre-breeding efforts have reported a benefit-cost-ratio with a substantial return of investments (Maredia et al, 2010). NordGen has initiated the planning of such public-private pre-breeding activities in the Nordic countries (NordGen, 2009; Nilsson and von Bothmer, 2010). Pre-breeding could prove to be an important activity where genebanks should take a leading role. The success of the wheat pre-breeding activities at CIMMYT under the leadership of Bent Skovmand could be seen as an example to study when developing the *best practices* for pre-breeding efforts (Dworkin, 2009).

The results from a survey reported in 1984 on the use of germplasm from genebanks among plant breeders in Europe can be read as a call for pre-breeding: *"There is a consensus of opinion that genebanks are not being used very extensively by breeders and certainly not nearly as much as they could be in relation to their potential value (...) breeders are likely to request samples only when they are needed for specific purposes. (...) The use of genebanks is far higher when linked closely to major breeding programmes"* (Peeters and Williams, 1984:22-23).

3.11 Seed requests: What do the genebank customers ask for?

In practice the genebank collection is made available for a wide variety of uses ranging wide in the number of requested samples. Based on the capacity of the germplasm evaluation program, the genebank customer usually request no more than a few hundred accessions, and often only one or a very few accessions. Most often the reported statement of the planned use describes a specific target. Only in a few cases the genebank customer is interested in a set representing the maximum diversity in the collection. Some of the intended use of requested germplasm materials reported with the seed requests received by the Nordic Genetic Resource Center (NordGen) during the previous year includes:

- *"Physiological tests for determination of Giberillic metabolism"*, Carlsberg Research Center (Denmark), 69 accessions
- *"Physiological tests for determination of Giberillic metabolism"*, Purdue University (USA), 100 accessions
- *"Request for material of Iranian origin"*, Iranian Biological Resource Center (Iran), 33 accessions
- *"I intend to use the seed material for research on the base broadening of Barley to create more sustainable crops of Barley"*, Scottish Agricultural College (United Kingdom), 3 accessions
- *"We would like to have the original donor of the Rfm1 gene in order to check molecular markers for Rfm1"*, Ackermann Saatzucht (Germany), 7 accessions
- *"I plan to screen these grain for the presence of fungal endophytes"*, AgResearch Grasslands (New Zealand), 27 accessions
- *"I will use the seeds in experiments testing different barley cultivars' tolerance against Bipolaris sorokiniana in different climate scenarios"*, University of Copenhagen, Department of Plant Biology and Biotechnology (Denmark), 1 accession
- *"Material is to be compared with material of the breeding company and for use of new characteristics in breeding program",* Bejo Seeds (Netherlands), 3 accessions

(Source: SESTO; http://sesto.nordgen.org/, verified 25 Jan 2011)

Plant breeders and other users of the Nordic genebank collection often ask for accessions with general description of the target use. Without relevant evaluation data available for the particular target trait, it is difficult for plant breeders to identify and select specific accessions. This implies that a *focused identification of germplasm strategy* (FIGS) can provide a useful service to assist plant breeders looking for relevant accessions. Even if there would be relevant documentation available for some accessions, for example for accessions in a Core collection set, selecting only from these accessions would miss out on the chance of finding new sources of desired genetic diversity. The Core collection has however been suggested as an approach to identify samples in the complete genebank collection that are similar to those observed to represent the desired characteristic in the core collection (Frankel and Brown, 1984 *cf*. Brown, 1995:6). This topic will be discussed further in more detail in chapter 6 and chapter 7.

On average each seed request at the NGB/NordGen during the last 20 years have resulted in the shipment of 12.6 accessions. During 2010 the average number of accessions shipped per request was 11.8, for 2009 the average was 16.5, and in 2008: 16.3 accessions (SESTO, http://sesto.nordgen.org/; Fredrik Ottosson, NordGen, personal communication).

3.12 Genebank funding in crisis

The establishment of a genebank is a long-term investment: *"Genebank work ought to be planned for a period of approximately 100 years. A shorter period of time could imply harmful discontinuity of the maintenance of the material (...) The local conclusion of this must be that the economical resources needed for maintenance of a sample for 100 years are anticipated when a sample enters the genebank"* (Yndgaard and Kjellqvist, 1984:34). Despite of the recognized key role genebanks have to ensure future food security, many of the genebank collections worldwide struggle with unstable funding and unstable political support (Flood, 2010).

3.13 Economic value of disease resistance research

Two Danish scientists compared in the 1980s the economic value of pest resistance research in barley to the economic gain that is achieved by fungicides in Denmark. They found from a large field experiment that barley powdery mildew research in Denmark contributed annually to a 3% actual reaped yield increase (Jørgensen and Kølster, 1985 *cf*. Bjørnstad, 2005). They further calculated that this corresponds to annual revenues of 260 million DKK (1985) compared to annual investments in the public and private sector estimated of 5 million DKK, and an estimated net revenue of 255 million DKK. The application of fungicides at a total cost of 250 million DKK in spraying costs contributed to a 305 million DKK value in yield increase, a net benefit of 55 million DKK. The comparative economic value of pest resistance research was thus shown to be substantial (Bjørnstad, 2005).

3.14 Do we need so many long-term genebank facilities?

Numerous other reports and scientific publications clearly indicate that the monetary benefits supported by the genetic material from genebank collections far exceed the costs of running these facilities - but few genebanks have sufficient and stable funds for the sustainable long-term conservation required to ensure that these collections of genetic resources remain available to future generations. The estimated 1700 genebanks established during the last 3 to 4 decades tell a tale of the perceived importance of such genebank facilities. However, the rapid increase in the number of genebanks may be part of the problem, rather than a solution.

The latest newsletter form the Crop Diversity Trust, Carry Fowler discussed the present large number of genebanks and suggested a possible similarity with the so-called *'bubbles'* in the stock markets (Fowler, 2010). In 1970 there was less than 10 genebanks for long-term storage of

germplasm, and today the *Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture* (FAO, 2010) provides an overview of more than 1700 genebanks, of which 130 are large genebanks defined as holding more than 10 000 accessions. This rapid increase in the number of genebanks justifies some reflections to be made. The 1989 annual report from IBPGR estimates the existence of 8 long-term genebanks in 1974, 55 in 1984, and 106 in 1990 (IBPGR, 1990). We further suspect that many of these new genebanks hold a substantial number of copies of germplasm samples stored at other genebanks. The new national genebanks often repatriate samples that have connections to their home country from the older and larger genebanks. The NGB/NordGen still today sometimes receives large unspecified seed requests for all germplasm material in the Nordic collection originating from the same country as the seed requestor. Fowler (2010) asked the obvious question: *"Can so many genebanks continue to persist and be supported in their present form without providing a clear verifiable service?"* Fowler also reminded us that each country is dependent on the sustainable conservation of germplasm from all parts of the world, and that this is a global problem calling for the creation of a global system. The larger international genebanks have the experience to ensure the long-term storage of germplasm, while the smaller national institutes have valuable local connections to the farmers and the local plant breeding industry and research institutes (Fowler, 2010). Would not a solution that takes better care of the particular strengths held by each participant be a better solution?

Already 30 years ago Frankel expressed himself with vigorous strength against the shift of focus from integrated global solutions to national genebanks: *"He urged greater use of the national genebanks (Frankel, 1987) and more comprehensive evaluation and documentation of accessions. He proposed the use of representative 'core collections' as being more accessible for plant breeders (Frankel, 1984). Nevertheless, Otto had always regarded the global network of base collections as the backbone of the genetic conservation strategy, and he was appalled when one of his erstwhile colleagues suggested a shift of emphasis to the national collections. At the age of 90 he still responded vigorously (Frankel, 1990)"* (Evans, 1999:177).

3.14.1 Duplication of accessions between genebanks

Holden (1984:282) presented estimates of involuntary duplications of genebank accessions between genebanks as high as 60%. Even if this estimate is more than 25 years old, there have been very few initiatives to identify duplicated accessions across genebanks. And when such studies have been performed, they have not always resulted in the verification and elimination of such duplicate accessions.

The Nordic Gene Bank [NordGen] classifies suspected duplicates of the same cultivars as a so-called *'temporary collection'* (NGB, 1987, 1990). This temporary collection does not undergo routine germination testing or regeneration. Samples with the temporary status are however published in the seed catalog and available for distribution. Only the sample considered as being the most original in a group of suspected duplicates are held under long-term storage (NGB, 1994:7; SESTO, http://www.nordgen.org/sesto, verified 25 Jan 2011). The European Integrated System (AEGIS) defines a so-called *'most appropriate accession'* with a similar function. The European crop working-groups identify with this concept the genetically unique and important germplasm samples held by genebanks in Europe (ECPGR, 2009:9,21; http://www.aegis.cgiar.org/, verified 25 Jan 2011).

Knüpffer (1989) developed an approach he called *'keyword in context'* (KWIK index) for the identification of possible duplicates across the European Barley genebank collections. Hintum and Knüpffer (1995) developed a terminology for the identification of duplicates between such genebank collections using the European Barley database and the KWIK index as an example.

Willner et al. (1998) classified duplicate accessions as *historical duplicate* when the accessions originate from the same collected germplasm or belong to the same source cultivar, and as *biological duplicate* when the accessions are demonstrated in an experiment to have the same genetic composition. At the ECPGR Forage Working Group meeting in 1999, the terminology *'Most Original Sample'* (MOS) was proposed together with a detailed algorithm (Maggioni et al., 2000:20-21, 214-217). The definition of MOS was developed further at the next Forage Working Group meeting (Boller et al., 2005).

Table 3.1: Estimated numbers for total world germplasm accessions for some of the crops with good representation in genebank collections

Crop	1984		1997	2010
Triticum (wheat)	410 000	(30%)	788 654	856 168
Oryza (rice)	215 000	(42%)	420 341	773 948
Hordeum (barley)	280 000	(20%)	486 724	466 531
Zea (maize)	100 000	(50%)	261 584	327 932
Phaseolus (phaseolus)	105 500	(38%)	268 369	261 963
Glycine (soybean)	100 000	(18%)	176 400	229 944
Sorghum (sorghum)	95 000	(32%)	168 550	235 688
Avena (oats)	37 000	(41%)	223 287	130 653
Arachis (groundnut)	34 000	(32%)	81 186	128 435
Lycopersicon (tomato)	32 000	(31%)	78 376	83 720
Pennisetum (pearl millet)	31 500	(49%)	36 806	65 447
Gossypium (cotton)	30 000	(27%)	48 889	104 780
Cicer (chickpea)	25 000	(54%)	69 736	98 313
Capsicum (capsicum)	23 000	(43%)	53 558	73 518
Saccharum (sugar cane)	23 000	(35%)	21 464	41 128
Vigna (cowpea)	20 000	(60%)	85 543	65 323
Secale (rye)	18 000	(44%)	27 132	21 192
Manihot (cassava)	14 000	(43%)	27 906	32 442
Lens (lentil)	13 500	(41%)	27 424	58 405
Allium (garlic, onion)	10 300	(49%)	25 288	29 898
Vicia (faba bean)	10 000	(50%)	31 831	43 695
Dioscorea (yam)	10 000	(50%)	11 500	15 903
Ipomoea (sweet potato)	8 000	(63%)	31 796	35 478
Beta (sugar beet)	5 000	(60%)	24 085	22 346

Source: Lyman, 1984 *cf.* Plucknett et al., 1987:111; FAO, 1997:92, 463-501; FAO, 2010:62, 244-283. For the values from 1984 the estimated proportion of distinct accessions is included.

The growth in number of genebank accessions worldwide still continue to increase as illustrated in table 3.1. Note however that it is not unlikely that the reported increase in the number of genebank accessions is in part the result of a more complete list of germplasm collections. However accessions are of course also added from new collecting expeditions and other sources (Lyman, 1984; Plucknett, 1987; FAO, 1997, 2010).

CHAPTER 4

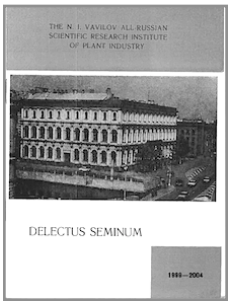
4. Genebank documentation

"The greatest constraint on utilization of plant genetic resources by researchers, taxonomists, breeders, farmers, and other users of germplasm presented in the strategies is the lack of accession level information, including passport, characterization (morphological and molecular), and evaluation data, especially for 'useful' traits. Even when data exists, low quality or reliability of characterization data, lack of use of standard descriptors, and constraints on access to information limit the usefulness of the data" (Khoury et al., 2010:11).

Documentation and associated information for germplasm samples preserved and made available from genebanks are essential for the rational use of these resources. This task has too often been neglected in the past as the frequently reported lack of even the most basic passport data for genebank accessions bear witness: *"The first SoW report highlighted the poor documentation available on much of the world's ex situ PGR. This problem continues to a substantial obstacle to the increased use of PGRFA in crop improvement and research"* (FAO, 2010:77). The development of appropriate routines and practices for the capturing of essential information during the acquisition of a new germplasm sample is the first step (1) to good genebank documentation. The development of a suitable genebank database management system to maintain and curate this information is the second step (2), and the focus of this chapter. The exchange and sharing of genebank data was the focus of PAPER I.

4.1 Short introduction to the history of germplasm documentation

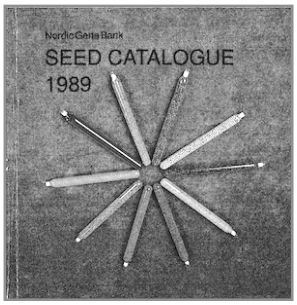
The botanical gardens established all over Europe starting from the 1540s were involved in the introduction of useful plants, including new agricultural crops. The *Index Seminum* [list of seeds] provided an early documentation and information sharing system and was distributed among botanical gardens, and beyond. The *Index Seminum* is a list of seed samples available for distribution. However, the seed samples were often the originally collected material or collected from the plants on display (open-pollinated regeneration) in the respective botanical garden. Often seeds with very low germination were distributed (Thompson, 1970). The production and publication of an *Index Seminum* was included as part of the mandate when the Nordic Gene Bank was established in 1979 (Palmstierna et al., 1975:48).



Delectus Seminum, 1999-2004, VIR (Dragavtsev et al., 1999)

Index Seminum Gaterlebensis, IPK Gatersleben (Knüpffer, 1999)

Türki Bitki Genetik Kaynaklari, *Index Seminum*, 1970-1980, (Settar, 1982)



Seed Catalogue 1989 (Nordic Gene Bank, 1989)

N. I. Vavilov established a large long-term collection of cultivated plants, including wild relatives. He demonstrated that the genetic diversity available in these genetic resources were the raw material for crop improvement (Hawkes, 1988). The genebank he established in Saint Petersburg is considered the first of its kind and the model for later genebanks (Frankel, 1988).

The Vavilov Institute published regularly a *Delectus Seminum* [list of selected seeds], a practice that later genebanks followed (PAPER I).

4.2 Types of germplasm data

Genebank documentation classifies the descriptors used to describe germplasm in a number of different categories. Passport data are the descriptors required for unique identification of a germplasm sample in a genebank collection. Passport terms include descriptors for the original collecting site or in the case of cultivated material, the pedigree. The description of the original collecting site includes country, place name, collecting date and geographic coordinates. Other passport terms describe the taxonomic classification and the identity of the genebank institute holding the germplasm sample. Management data describe routine germination and viability tests and the location and amounts held in the seedstore. Other important management descriptors keep track of the conditions and status of germplasm regeneration cycles. The genebanks hold collections of living material. The seeds require multiplication when the number of seeds is getting too low for distribution and rejuvenation when the germination and vitality is falling below defined levels. These data are important to assess the health and genetic integrity of the conserved germplasm material.

- Passport data (identity, origin)
- Management data (collection logistics)
- *In situ* population data (for *in situ* conservation)
- Characterization and evaluation data (C&E)
    - Characterization data (phenotype stable across environments)
    - Evaluation data (phenotype influenced by environment)
    - Molecular data (genotype and chemical analysis)

(Source: IPGRI, 1994)

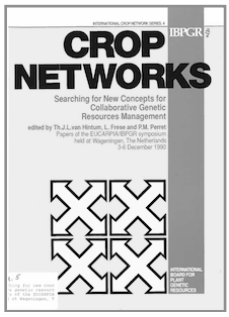
Perhaps more important to support the use of plant genetic resources are the so-called characterization and evaluation (C&E) descriptors. Trait properties useful for plant breeders and other users of the material are described by these terms. These descriptors are often divided further as those phenotypic characters (characterization data) that are stable across different environments and thus useful for the identification of distinct cultivars, landraces or sometimes even wild populations; and those that show an interaction with the environment (evaluation data). Genebank curators use characterization data to assess and avoid genetic drift. Evaluation data include trait properties that are influenced by the environment the plants are exposed to. This is also called a genotype by environment interaction (GxE) effect. Evaluation data include disease and pest resistance of particular interest to plant breeders. A last category belonging to the characterization data is the molecular marker data and other similar information derived from molecular studies of germplasm.

4.3 Computerized genebank databases and information networks

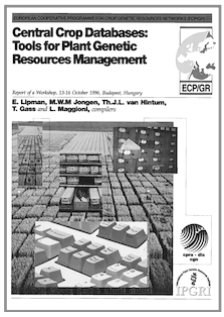
One of the first initiatives to develop international standards and data exchange mechanisms for sharing computerized information on crop genetic resources was presented by the Food and Agriculture Organization of the United Nations (FAO) and the International Atom Energy Agency (IAEA) at the *'Fifth Yugoslav Symposium on Research in Wheat'* in 1966 (Konzak and Sigurbjörnsson, 1966). The proposal from the assembled expert group was to establish a distributed genebank information network with a central hub and a central database hosted by FAO in Rome. The development of data standards and distributed information networks for plant genetic resources are covered in more detail in PAPER I.

One of the first electronic information systems developed for documentation of genebank material was the EXIR (Executive Information Retrieval) system (Hersh and Rogers, 1975). EXIR is based on the TAXIR (Taxonomic Information Retrieval) genebank information system developed by the Taximetrics Laboratory at the University of Colorado, USA (Estabrook and Brill, 1969). The IBPGR (International Board for Plant Genetic Resources) established in 1974 was from the beginning involved in the further development of the EXIR system for documentation of genebank data (IBPGR, 1976). Work was also initiated on a distributed genebank information system under the name GR/CIDS (Genetic Resources Communication, Information and Documentation System: *"When it is fully developed, GR/CIDS should encompass the whole of the information component, including documentation and the flow of information, of genetic resources work, from the initial collection of data about traditional materials in the field to the performance of improved varieties derived from them"* (IBPGR, 1976:5). It is possible that the GR/CIDS did not continue to develop into a true distributed and long-term global system based on the inherent expectation of a strong standardization of information services from the national genebanks and other target data providers.

European Catalogue of Genus Secale First Edition (Serwiński and Konopka, 1984)



Crop Networks (Hintum et al., 1991) EUCARPIA / IBPGR



Central Crop Databases, Tools for PGR Management (Lipman et al., 1997)

EURISCO, European Search Catalogue for Plant Genetic Resources, http://eurisco.ecpgr.org

The genebank documentation network in Europe was coordinated through the European Cooperative Programme for Plant Genetic Resources (ECPGR). In the 1980s and 1990s, a number of European Central Crop Databases (ECCDBs) were established. The first ECCDB was the Rye Catalogue, compiled by the Polish Gene Bank between 1981 and 1984 (Serwiński and Konopka, 1984). The Nordic Gene Bank compiled the European Prunus Database (EPDB) published as a dBase database in 1984 (Kurki, 1986). In 1989, NGB published a series of printed catalogues with the ECCDBs of cherry, apricot, almond, peach and plum (Bjarnason and Niklasson, 1989). The ECCDBs for *Phleum*, *Agrostis*, and *Phalaris* were published online by NGB in 1996 and 1997 (Huldén, 1997). There are currently 58 ECCDBs under the platform of the ECPGR coordinated from a total of 6 crop networks and 18 crop-based working groups (http://www.ecpgr.cgiar.org/Databases/). These crop networks provide important coordination of joint European activities on conservation and use of plant genetic resources in Europe.

4.4 EURISCO, European Search Catalogue for Plant Genetic Resources

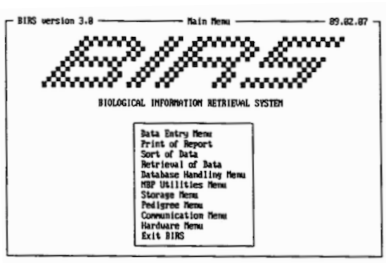
The EU-funded EPGRIS project (Establishment of a European Plant Genetic Resources Information Infra-Structure; 2000-2003) (IPGRI, 2001, 2002) produced the European Search Catalogue for Plant Genetic Resources (EURISCO, http://eurisco.ecpgr.org, verified 25 Jan 2011). Theo van Hintum representing the Dutch genebank in Wageningen coordinated the EPGRIS project. The final product, the EURISCO database, was released in 2003 (EURISCO, 2003; IPGRI, 2003) and hosted at IPGRI in Rome, Italy (IPGRI was reorganized as Bioversity International in 2007). After the initial project phase, Bioversity International hosted the EURISCO portal with support provided from the ECPGR Documentation and Information Network (Maggioni, 2005, 2007, 2010).

The EURISCO database is today considered one of the three most important data portals for plant genetic resources, the other two being SINGER (the System-wide Information Network for Genetic Resources; http://singer.cgiar.org/, verified 25 Jan 2011; Ninnes et al., 2002) of the CGIAR (Consultative Group on International Agricultural Research); and the USDA NPGS GRIN (United States Department of Agriculture, National Plant Germplasm System, Germplasm Resources Information Network; http://www.ars-grin.gov/npgs/, verified 25 Jan 2011). The global accession level germplasm system, GeneSys (Gateway to Genetic Resources; http://www.genesys-pgr.org/, verified 25 Jan 2011) will foremost build on the EURISCO, SINGER and USDA GRIN to create an integrated information system for these three distributed platforms. Information from other genebank datasets will also be included (FAO and ITPGRFA, 2008).

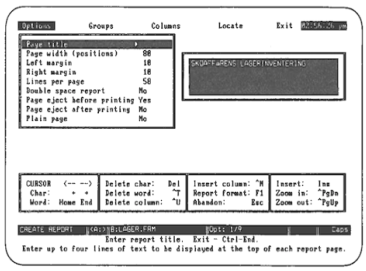
4.5 Documentation system at the Nordic Gene Bank, 1979-1989

The Weibullsholm *Pisum* collection assembled by Herbert Lamprecht (1889-1969) and Stig Blixt (1930-2009) can in many aspects be seen as the predecessor to the Nordic Gene Bank (NGB, 1991; Ellerström, 1982; Yndgaard and Kjellqvist, 1982). The Weibullsholm *Pisum* database maintained by Stig Blixt became the most important model for the first genebank database at the Nordic Gene Bank (Yndgaard, 1982). First an evaluation of alternative solutions was made (Lindeberg et al., 1981). The *Pisum* system was here compared to other systems such as the EXIR (Executive Information Retrieval) system (Hersh and Rogers, 1975). However the requirement of expensive hardware and the list of required modifications including those for the data model to meet the needs specified by NGB, made the evaluation report recommend the NGB information system to be based on the *Pisum* database (Lindeberg et al., 1981).

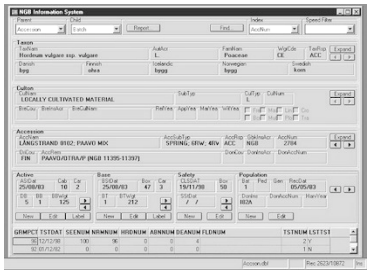
The first database management system for the Nordic Gene Bank was developed internally by NGB and named Biological Information Retrieval System (BIRS). Between 1983 and 1988 BIRS was in use on the IBM minicomputer platform, and later recompiled for the IBM PC when this platform became available (Rydström, 1989). In 1989 the BIRS system was migrated as a module to the NOBIS (Nordic Biometry System; Yndgaard, 1990) platform of the Nordic Biometry Project (NBP). BIRS and NOBIS were distributed and used by crop breeders and researchers in the Nordic region. BIRS and NOBIS provide thus an early example of a distributed data network for documentation and exchange of standardized NOBIS format data files with evaluation (C&E) data from experiments across all the five Nordic countries. Data exchange was done by floppy diskettes or by modem over the public phone line.



Biological Information Retrieval System, BIRS (1983-1989)



Borland dBASE III, and IV (1989-1995)

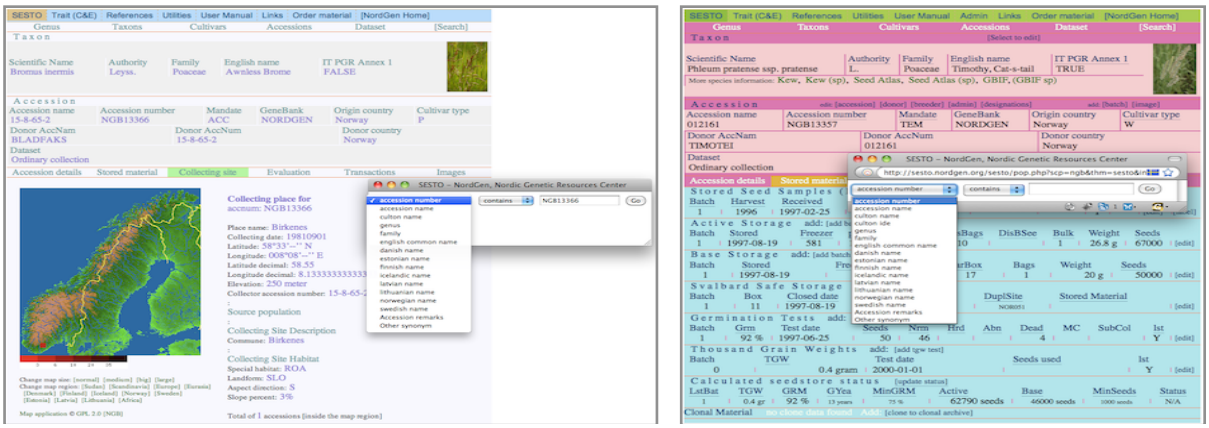


NGB Information System (Huldén, 1995-2001)

4.6 Documentation system at the Nordic Gene Bank, 1990-2002

During the 1990s the NGB documentation system was gradually migrated to the Borland dBase IV database format, and in 1995 to the Visual dBase with the development of the first version of SESTO (SEedSTOre) user interface (Huldén et al., 1998). NGB started in 1994 to publish an online search interface to the genebank database using a text-indexing engine called 'freeWAIS'

and the Apache web server on a Linux system (Huldén, 1999). This initiated the move towards open source software at the Nordic Gene Bank.

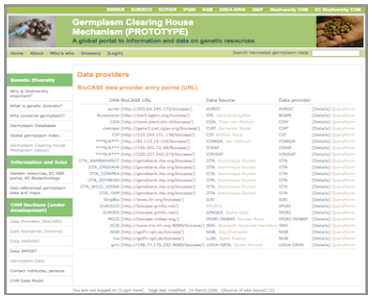


SESTO, genebank management system (NGB, NordGen, 2002-2011) PHP version, http://sesto.nordgen.org/sesto/ (visited November 3, 2010).

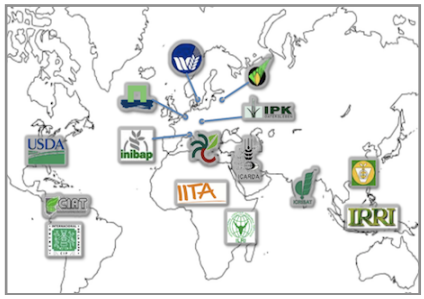
4.7 SESTO genebank management system for the Nordic countries

In 2002 NGB developed and released a new and online version of SESTO. The new version was developed using the PHP (Hypertext Preprocessor) computer scripting language and the PostgreSQL database system (Endresen, 2003; Endresen et al., 2005a). The NGB staff could authenticate themselves using a personal login to get access to update the database from the online web interface. This PHP version of SESTO was developed as a generic framework with the capacity to host the genebank database for multiple genebanks (Endresen et al, 2005a, Endresen et al, 2005b). Parts of this PHP version of SESTO were installed at the European Barley Database (ECCDB; Knüpffer, 1988) hosted by IPK Gatersleben in Germany, primary for the presentation of C&E trait data (http://barley.ipk-gatersleben.de/genres/, verified 25 Jan 2011) (Enneking et al., 2003:16). A demo version of SESTO was in May 2003 installed at VIR in St Petersburg. SESTO was fully implemented and is used by the Baltic genebanks in Estonia, Latvia and Lithuania since around 2004 (JPBI, 2005). The documentation manager from the new genebank in Bhutan received training at Alnarp with SESTO for a few weeks in October 2003. The NordGen SESTO system was also presented for the SEEDNet genebanks of Southeast Europe, and installed in Albania. In September 2007 NGB received Kim Kyang from North Korea for SESTO training. In East Africa training with SESTO was provided for the EAPGREN genebanks of the ASARECA region. Johan Bäckman and Magdalena Svärdh visited Uganda in November 2007 and Sudan in April 2008 for on site installation of SESTO at the respective national genebanks. Training was also provided with SESTO for the genebank network in Central Asia and Caucasus (CAC). Jonas Nordling and Simon Jeppson from NordGen installed SESTO in Kyrgyzstan and Tajikistan in December 2010.

GCP Central Registry (2005) http://gcpcr.grinfo.net



Germplasm CHM demo (2006) http://chm.grinfo.net



BioCASE installation at genebanks during 2005 and 2006.

4.8 Generation Challenge Program, Central Registry (GCPCR)

During 2005 a new version of the Generation Challenge Program, Central Registry (GCPCR) was developed by Tom Hazekamp (IPGRI) and Dag Endresen in the role as consultant for IPGRI. The GCPCR is a repository for research data generated by the respective research projects of the GCP. The computer code of this new GCPCR system was based on the source code from SESTO. This new version of the GCPCR was launched at the GCP Annual Research Meeting in 2006 (http://gcpcr.grinfo.net, verified 25 Jan 2011). The GCPCR was later developed further at Bioversity (Skofic et al., 2009). One important outcome from this software development the start of a general framework upon later germplasm data portals successfully was developed. Some of these derived germplasm portals are presented below.

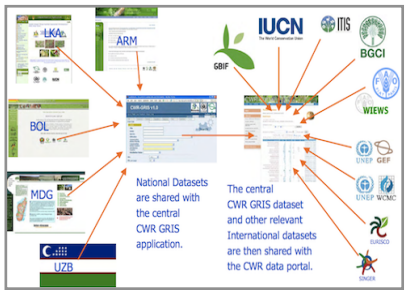
4.9 Germplasm Clearing House Demo Portal

In 2006 a demo portal was developed for a distributed genebank information network based on the standards and software tools maintained by Biodiversity Standards (TDWG) and the Global Biodiversity Information Facility (GBIF). This demo portal was developed using the computer code from the GCPCR and further refined the features of these lines of computer code as a general customizable platform for similar information systems. The working title for this demo project was *'Germplasm Clearing House Mechanism, A global portal to information and data on genetic resources'* (http://chm.grinfo.net, verified 25 Jan 2011). Walter Berendsohn (Berlin Botanical Garden) and Helmut Knüpffer (IPK Gatersleben) provided the mapping between the central data standard used by the BioCASE data publishing toolkit (Biological Collection Access Service for Europe; http://www.biocase.org, verified 1 Feb 2011) and the Multi-Crop Passport Descriptors (MCPD; Hazekamp et al., 1997; Alercia et al., 2001) used by the genebank community (Berendsohn and Knüpffer, 2006). This mapping opened up possibilities for using the BioCASE software to build a distributed information network for the genebank community (Dag Endresen and Javier de la Torre, unpublished results; ABSTRACT 4). The BioCASE toolkit was installed at most of the CGIAR centers, the USDA GRIN system in America, and at some selected genebanks in Europe (Germany, Netherlands, Latvia and the Nordic countries) (ABSTRACT 2, and ABSTRACT 3). Because of the interoperability with the BioCASE standards the genebanks included in this demo project could also connect to the distributed biodiversity information network established by the Global Biodiversity Information Facility (GBIF, http://data.gbif.org, verified 25 Jan 2011) located in Copenhagen, Denmark (Knüpffer et al., 2007; ABSTRACT 3).

4.10 Crop Wild Relatives Information Network

In the autumn of 2006 the development of a new information system with focus on conservation for the wild relatives of the cultivated plants was initiated at IPGRI (IPGRI was reorganized as Bioversity in 2007). A few years earlier a EU-funded project called *'PGR Forum'* (http://pgrforum.org, verified 25 Jan 2011), coordinated by Nigel Maxted (England) produced a similar information system, CWRIS (Crop Wild Relative Information System) (Moore, et al., 2008). The new CWR information system IPGRI/Bioversity started to build in 2006 was not built on the CWRIS, but was meant to build on another information system developed at FAO called *'GeoNetwork'*. However the GeoNetwork system was designed for the sharing of spatial dataset and resources, and the required modifications for the needs of the Crop Wild Relatives Portal proved to be too many. Instead some modifications of the information system developed for the GCPCR and the CHM demo portal were made on relative short notice, and the first version of the Crop Wild Relatives Global Portal (http://www.cropwildrelatives.org, verified 25 Jan 2011) was ready for release as a prototype demo portal including CWR data from the first 5 countries in August 2007 (Thormann et al., 2007). The CWRIS as well as the later CWR information system included components to manage passport and *in situ* population data.

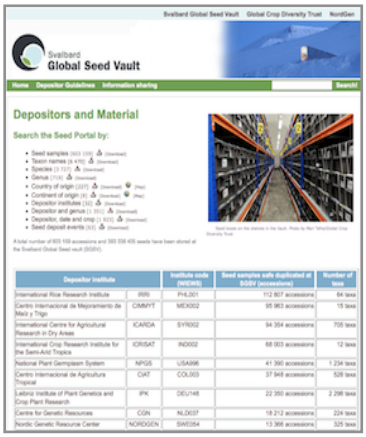
PGR Forum (2003-2005)
http://pgrforum.org

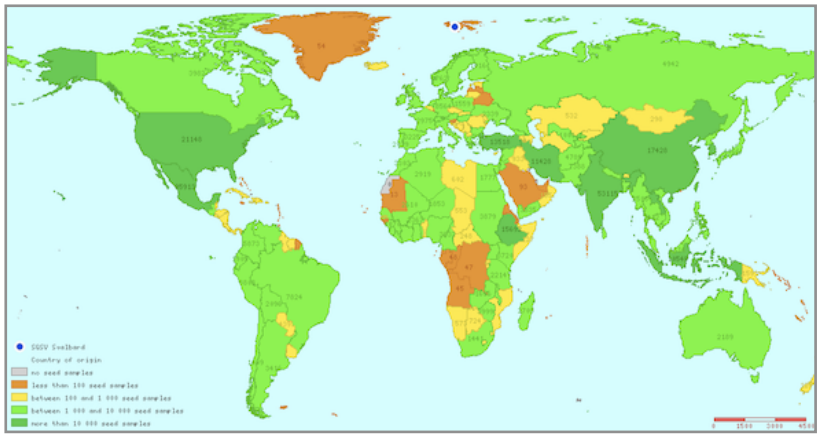


Crop Wild Relatives Information Network (2007-2011)
http://www.cropwildrelatives.org

4.11 Svalbard Global Seed Vault Data Portal

The computer code developed as generic framework for germplasm information systems proved in 2007 to be useful for another major data portal for plant genetic resources. In 2004 a second proposal for international seed storage at Svalbard was initiated. During the feasibility study, the need for a documentation system came up, NGB suggested to use the SESTO genebank management system (http://sesto.nordgen.org/sesto/index.php?scp=svalbard, verified 25 Jan 2011; Endresen et al., unpublished report). However the Seed Vault was designed as a safety deposit repository and very different from a standard genebank. There was thus no need for the full complexity of the SESTO system. Instead a new portal for the Svalbard Seed Vault was developed based on the generic portal software that was mentioned above. [Personal comment: I remember sitting on the ice-cold floor inside the permafrost vault the evening before the official opening to complete the final portal to be ready for the inauguration.] The final Seed Vault portal (http://www.nordgen.org/sgsv/, verified 25 Jan 2011) was released for the opening on 26 February 2008 and publishes literary all the information the NGB receives from the seed depositors openly to the Internet. Johan Bäckman (NGB) developed a separate logistics database with information about the precise location of each depositor seed box inside the Vault. These tables are not connected to the public Data Portal for the Seed Vault. Ola Westengen in his role as the Coordinator of operation and management for the Seed Vault is responsible for updating this logistics system during his visits to the Vault. The master copies of the Seed Vault databases are located at NordGen in Alnarp (and securely backed up elsewhere) there is also a paper printout with all data deposited inside the Vault.





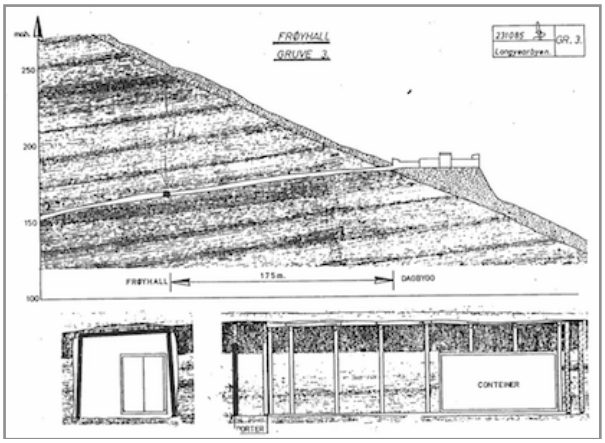
Screenshots from the data portal of Svalbard Global Seed Vault, http://www.nordgen.org/sgsv/

4.11.1 Nordic Gene Bank, *Frøyhallen* (1984)

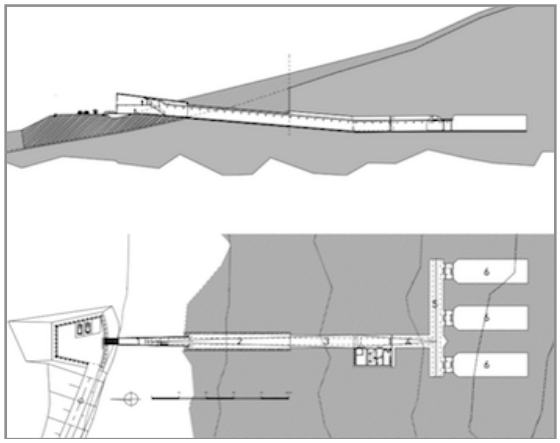
The Nordic Gene Bank deposited on 14 November 1984 the first 203 seed samples at Svalbard. The facility was named *'Frøyhallen'*. Arne Wold and Flemming Yndgaard carried together the first seed box into *Frøyhallen*. The original initiative for the Svalbard location came from Arne Wold at the Seed Testing Agency in Norway in a letter to NGB dated 3 November 1982. The

storage room was a small vault in an abandoned coalmine, 4 m wide, 20 m deep, with 3 m height under the roof, and located approximately 175 m inside the mountain. Inside the vault was placed a metal container, 1.6 m wide, 4 m deep, and 2 m height. The permafrost conditions provided natural cooling. For this storage facility NGB entered a leasing contract for the payment of approximately 6000 NOK each year, plus 1000 NOK as a one-time handling fee for each seed box (holding 400-500 accessions). The exact amounts for the leasing costs were edited out of the final version of the agreement (SNSK and NGB, 1984). The total costs of preparing the vault, including the cost of the metal container was in July 1984 estimated to 120 000 NOK. The natural permafrost inside the vault was measured at a stable -3.6 °C to -3.8 °C. (Source: NGB correspondence archive; Yndgaard, 1983:30, 1985; Jadav et al., 2003).

The concept of permafrost seed storage attracted immediately international attention. The IBPGR organized a workshop at Reading, UK in April 1985 with the title *'Cost-Effective Long-term Seed Stores'* where the use of natural facilities located in natural caves, high altitude sites, permafrost, glaciers, and the Antarctica was discussed (IBPGR, 1985:12-13). Flemming Yndgaard presented the NGB Svalbard facility (IBPGR, 1985:31). Yndgaard also published the same year a manuscript with a more detailed description of the NGB Svalbard facility in the *Plant Genetic Resources Newsletter* (Yndgaard, 1985).



Svalbard, Frøys Hall, 1984. Drawing by Store Norske Spitsbergen Kullkompani (SNSK)


Svalbard Global Seed Vault; opened 25 February 2008. Drawing by Statsbygg (2008:17)

4.11.2 Svalbard International Seedbank (1989)

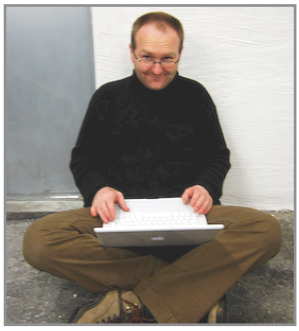
In July 1988 the International Board for Plant Genetic Resources (IBPGR) made contact with Arne Wold at the Norwegian Seed testing Agency regarding the possibilities for an international safe-duplication storage facility at Svalbard. One year later, during 11 - 13 September 1989 Y.J. Adham from IBPGR and R. Smith from Kew Garden, UK visited the NGB safe storage, *Frøyhallen*, together with Arne Wold. The report prepared by Smith (1989) concluded that: *"the choice of Spitzbergen, Svalbard, cannot be bettered"*. Despite the positive response from Sysselmannen at Svalbard and from the Norwegian Government the *'Svalbard International Seedbank Expert Consultation Meeting'* convened by FAO in 1990 found a few major barriers: (1) The political and legal status of plant genetic resources as common human heritage and a free good or alternatively as national sovereign property was under heavy debate; (2) the permafrost condition ensured only -3.5 °C while the international standard for rational long-term seed storage was prescribed as -18 to -20 °C (Cromarty et al, 1982; Smith et al., 2003); (3) the presence of Russian settlements at Svalbard in the time of the Cold War; and (4) the lack of long-term financial commitments for the operation of the facility (Fowler et al., 2004). The development of the *'Svalbard International Seedbank'* (SIS) concept at FAO and IBPGR/IPGRI continued for many years until it was finally abandoned around 1995. Qvenild (2005, 2006) suggested that the major obstacle was the lack of a legal framework for plant genetic resources

and the lack of trust between the developing countries and the developed countries. Even if the International Undertaking (FAO, 1983) established concepts such as *Farmers' Rights* and defined PGR as *'common heritage of mankind'*, this agreement was non-binding, and in other fora the frameworks for *Breeders' Rights* were developing fast.

4.11.3 Svalbard Global Seed Vault (2008)

In 2004 the Cold War was over, and the lengthy negotiations on the legal status for plant genetic resources resulted first in the Global Plan of Action, GPA (FAO, 1996b) and finally in the legally binding International Treaty on Plant Genetic Resources for Food and Agriculture, ITPGRFA (FAO, 2002). Carry Fowler took a central part of both the GPA and the ITPGRFA negotiations. However it was Henry Shands who brought the Svalbard idea "back to life" in an email to Carry Fowler dated 8 August 2003 (Qvenild, 2005, 2006). In March 2004, Emile Frison representing IPGRI and CGIAR sent a letter to the Government of Norway with a request too re-open the proposal on an international facility for seed storage at Svalbard. In May 2004, Noragric (Center for International Environment and Development Studies) received a positive response from the Norwegian Ministries with approved funding to conduct a feasibility study together with the Nordic Gene Bank (Fowler et al., 2004). A technical report prepared by Geoffrey Hawtin, former director of IPGRI (Hawtin, 2004) prepared for the approval of this second proposal for the Svalbard facility at the FAO Commission meeting in November 2004 (Qvenild, 2005, 2006). In the spring of 2007 Statsbygg in Norway started the construction of the Vault. The first seed box for the official opening on 26 February 2008 was carried into the seed vault by the president of the European Commission José Manuel Durão Barroso and the Norwegian Prime Minister Jens Stoltenberg.

Inside the Seed Vault, Ola Westengen, Johan Bäckman and Simon Jeppson. Photo by Dag Endresen, 27 Feb 2008



Dag Endresen in the Seed Vault. Photo Feb 2008 by Simon Jeppson

Svalbard Global Seed Vault, February 2008 by Dag Endresen

4.12 Integration of genebank information to biodiversity networks

Following in the footsteps of the first Botanical Gardens the Nordic Gene Bank (NGB), the Polish genebank (IHAR, Radzików) and the German genebank (IPK Gatersleben) were the first *ex situ* genebanks to join the Global Biodiversity Information Facility (GBIF) in 2004 (Knüpffer et al., 2004). As mentioned in 4.9, the first round with installation of GBIF compatible data publishing software for genebanks during 2005 and 2006 was with a focus on the CGIAR genebanks (ABSTRACT 2; Arnaud et al., 2008). In 2009 the GBIF secretariat released the first prototype versions of a new data publishing toolkit for biodiversity collections dataset. This software application was called GBIF Integrated Publishing Toolkit (IPT) was developed with the Java programming language (http://code.google.com/p/gbif-providertoolkit, verified 25 Jan 2011). The Java programming language is more commonly used in the PGR community than the Python scripting language of the previous BioCASE provider.

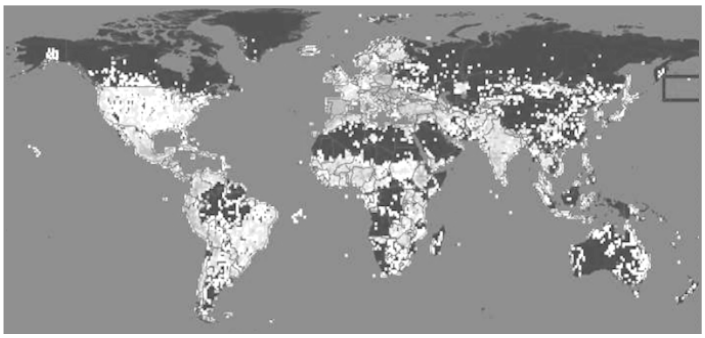
4.12.1 Darwin Core extension for genebanks

The IPT is based on the *Darwin Core* biodiversity standard (TDWG, 2009). The BioCASE software implemented another similar biodiversity standard called *'Access to Biological Collection Data'* (ABCD) (TDWG, 2005). The BioCASE software required a careful mapping of the terms from the ABCD standard to the genebank standards as well as a few additional terms before being useful for the sharing of genebank datasets (ABSTRACT 2, 3). For these same reasons the Darwin Core standard also required the mapping of terms and also some additional new terms before the IPT software that is based on this standard, was useful for the genebanks (ABSTRACT 5). The development of an extension to the Darwin Core standard resulted in the *'Darwin Core extension for genebanks'* (DwC germplasm; http://rs.nordgen.org/dwc/, verified 25 Jan 2011) as is described with ABSTRACT 5 and PAPER I.

4.13 EURISCO demo project for web services funded by GBIF

The implementation of the DwC extension for genebanks in the IPT software was made without any significant problems (PAPER I) and resulted in a new round of installations in the genebank community of this new GBIF compatible software. GBIF made available seed money for the installation and testing of the prototype version of the IPT software. NordGen was approved funding for a project in collaboration with the EURISCO team at Bioversity International to install the prototype IPT software for genebanks in Europe during 2010 (ABSTRACT 7).

Installations of the GBIF Integrated Publishing Toolkit (IPT) for genebank in Europe made during 2010.



GBIF Network 2 for plant genetic resources, http://data.gbif.org/datasets/network/2

4.14 Conclusion

The lack of available relevant information on the germplasm material provided by the genebanks remains a major constraint to their use (Khoury et al., 2010). The access to the available information is limited by the lack of interoperability between the major germplasm catalogues and between the online information systems provided by individual genebanks (FAO, 2010). Many of the older germplasm accessions also lack basic passport data recorded even in non-digital forms. This thesis contributes with new solutions for the interoperability and exchange of germplasm data between genebank information systems (PAPER I). The interoperability with other relevant biodiversity communities are also supported by using established standards from biodiversity informatics and contributing to the further development of these standards. Conservation of biodiversity including agrodiversity and the plant genetic resources is increasingly a cross-sector activity. The genebanks are more often required to make information on plant genetic resources available for new uses and new user-groups coming from policy and decision-making as well as other scientific research communities. The value of the efforts to improve the cross-sector interoperability of genebank information is thus expected to receive increased value and attention. This thesis contributes to such solutions.

CHAPTER 5

5. Ecogeographic data analysis

"Systematic botany has to document, present descriptively and classify in an organized manner the tremendous global plant diversity. This is a practical necessity: Mankind needs orientation and means for efficient communication about diversity. Only if based on this, are rational use and scientific investigation of plants possible, and these have to establish the foundation for continuous improvement of plant use which is necessary for the increasing human population on earth" (Mansfeld, 1962; translated from German by K. Pistrick *cf*. Pistrick, 2003:30).

This chapter discusses the ecogeographic data analysis as a pre-requisite for efficient conservation and use of plant genetic resources. The ecogeographic analysis would naturally include the assessment of taxonomic nomenclature concerning the entire genepool for the crop species in question. The controversy concerning the naming of cultivated species dates back to Linnaeus (Linnaeus, 1737; Stearn, 1986), but is still active in present day taxonomy. The concepts behind ecogeographic "gap analysis" are also discussed. Ecogeographic gap analysis is no modern "invention", however, the tools to undertake this type of gap analysis have been significantly refined in the recent years. Today powerful ecological niche modeling algorithms are used to identify insufficiently sampled genetic diversity for the planning of targeted collecting missions to fill "gaps" in the genebank collections (Ramírez-Villegas et al., 2010). The purpose of "gap analysis" is to find the most efficient collecting strategy to increase the total genetic diversity held in *ex situ* genebank collections. Similar species distribution-modeling tools can also be used to improve the assessment and optimal planning of genetic reserves (Maxted et al., 2008), and to assess the potential distribution and risk areas for invasive crop pests (Ganeshaiah et al., 2003).

5.1 Flora of Cultivated Plants in the USSR

During the first years of the Bureau of Applied Botany (predecessor of the present N.I. Vavilov Research Institute of Plant Industry) some of the staff members published ecogeographic studies such as: *"On the development of wild and weedy oats"* (Maltzev, 1914), *"Identification of wheats"* (Flyaksberger, 1915a), *"Wheat varieties in Siberia"* (Flyaksberger, 1915b), and *"On the origin of the cultivated rye"* (Vavilov, 1917). These early ecogeographic studies initiated further similar studies on the locally adapted ecotypes of cultivated species, and lead eventually to a longer series of comprehensive ecogeographic surveys. Evgenii Vladimirovich Wulff (1885-1941) was the series editor for the *"Flora of Cultivated Plants in the USSR"* (Wulff, 1935*), and Konstantin Andreevich Flyaksberger (1880-1939) was the editor of the first volume on wheat (Flyaksberger, 1935). Probably a coincidence, but this was only one year after the publication of the first volume of Komarovs Flora of the USSR (Komarov, 1934). During the next five years, the Vavilov Institute published more than 20 volumes of the *"Flora of Cultivated Plants"*, and some of them in multiple parts such as for the second volume of grain cereals (part 1 on rye, part 2 on barley, and part 3 on oat). An overview of the different volumes of the Flora of Cultivated Plants is provided on the Vavilov Institute home page, list of references at http://www.vir.nw.ru/books/vir_publ.htm (verified 25 Jan 2011). The *Flora of Cultivated Plants* comprises 22 volumes (Loskutov, 1999:35).

Kulturnaya Flora SSSR [Cultivated Plants of the USSR]

Vol. I: Wheat

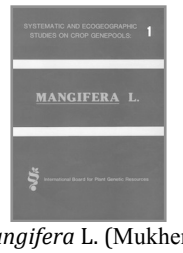
Vol. II Part 1: Rye

Vol. II Part 2: Barley

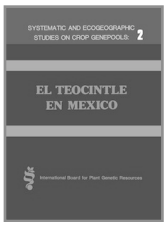

Evgenii V. Wulff (1885-1941), editor

5.2 Ecogeographic surveys of crops (IBPGR, IPGRI)

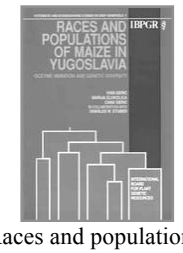
Bioversity International (previously IBPGR, 1974-1991; IPGRI, 1991-2006) is publishing a series of *'Systematic and Ecogeographic Studies on Crop Genepools'*. The series includes at the present 12 handbooks with a detailed description of the taxonomy, genetic diversity, geographic distribution, ecological adaption and ethnobotany of the respective plant groups (Mukherjee, 1985; Sánchez and Ordaz, 1987; Geric et al., 1989; Edmonds, 1990; Nabhan, 1990; Edmonds, 1991; von Bothmer et al., 1995; Maxted, 1995; Lira Saade, 1996; Hijmans et al., 2002; Maxted et al., 2004; Baudoin et al., 2004).



1. *Mangifera* L. (Mukherjee, 1985)

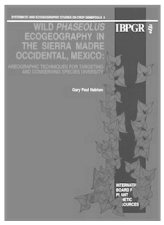


2. El Teocintle en México (Sánchez and Ordaz, 1987)

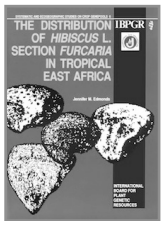


3. Races and populations of maize in Yugoslavia (Geric et al., 1989)

4. Herbarium survey of African *Corchorus* L. species (Edmonds, 1990)



5. Wild *Phaseolus* Ecogeography in the Sierra Madre Occidental, Mexico (Nabhan, 1990)



6. The distribution of *Hibiscus* L. section Furcaria in tropical East Africa (Edmonds, 1991)

7. An ecogeographic study of the genus *Hordeum*, 2nd edition (von Bothmer et al., 1995)

8. An ecogeographical study of *Vicia* subgenus *Vicia* (Maxted, 1995).

9. Estudios taxonómicos y ecogeográficos de las Cucurbitaceae latinoamericanas de importancia económica (Lira Saade, 1996)

10. Atlas of Wild Potatoes (Hijmans et al., 2002)

11. An ecogeographic study African *Vigna* (Maxted et al., 2004)

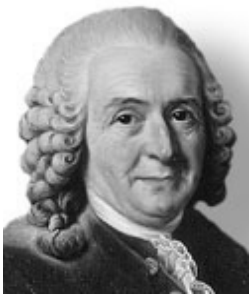
12. Ecogeography, demography, diversity and conservation of Phaseolus lunatus L. in the Central Valley of Costa Rica (Baudoin et al., 2004)

A task group of 12 persons formed by the International Board for Plant Genetic Resources (IBPGR) developed, in 1984, new guidelines for ecogeographic surveying and *in situ* conservation at a meeting held at Beltsville, Washington DC (IBPGR, 1985). The most recent guidelines for ecogeographic surveys were developed by a team of experts and published by

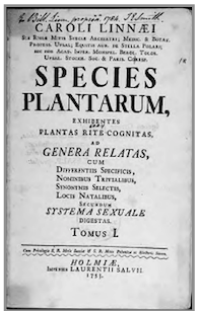
IPGRI in 2005 (Guarino et al., 2005). Maxted et al. (1997, 2008) developed more recent guidelines for *in situ* conservation strategies. Brush (2000) and Jarvis et al. (2007) developed comparable on-farm management routines.

5.3 The naming of cultivated plants

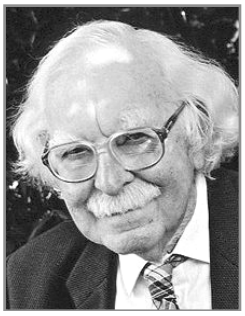
The taxonomy of cultivated plants differs from that of wild plants in significant ways. The cultivated forms follow migrating people and can be shared by exchange of culture and technology when people makes contact with each other. Cultivated plants are also under severe selection pressures exerted by humans through active selective breeding or through the adaptation to the agro-ecological environment.



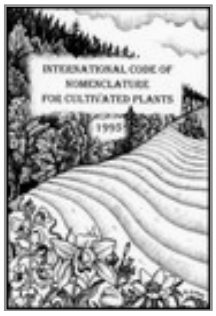
Carl Linnaeus
(1707-1778)



Species Plantarum
(Linnaeus, 1753)



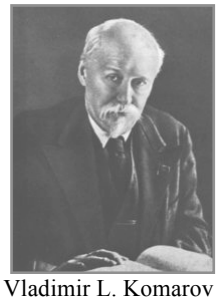
William Thomas Stearn
(1911-2001)



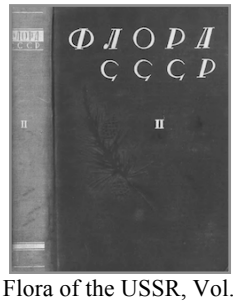
Cultivated Plant Code,
ICNCP (Stearn, 1953)

5.3.1 Carl von Linnaeus on the cultivated flora

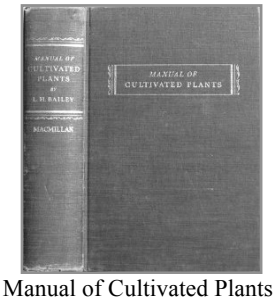
Carl von Linnaeus (1707-1778) can be considered as the *'father'* of modern botanical nomenclature with *'Species Plantarum'* from 1753 (Linnaeus, 1753). Linnaeus strongly opposed to dealing with systematic names for cultivated plants (in 1737 when he was 30 years old): *"All monstrous flowers and plants derive their origin from natural forms. (...) I distinguish the species of the Almighty Creator, which are true from the abnormal variation of the Gardener. (...) Such monstrosities, variegated, multiplied, double, proliferous, gigantic, wax fat and charm the eye of the beholder with protean variety so long as gardeners perform daily sacrifice to their idol. (...) Let this things continue, if it is desired, in the realm of gardening, but let us banish from sound systematic botany all plants with flowers multiplied or double or proliferous"* (Linnaeus, 1737:188-189 (aphorism 271), translated by Stearn, 1986:21). This dogmatic division impacts taxonomic research up to the present.



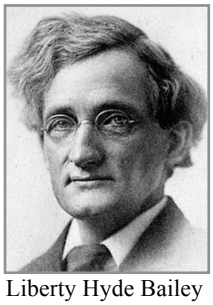
Vladimir L. Komarov
(1869-1945)



Flora of the USSR, Vol.
II (Komarov, 1934)



Manual of Cultivated Plants
(Bailey, 1924, 1949)



Liberty Hyde Bailey
(1858-1954)

5.3.2 Komarov's flora of the USSR

Vladimir Leontyevich Komarov (1869-1945) was the senior editor of the Flora of the USSR, published in 30 volumes between 1934 and 1960 (Komarov, 1934, 1968; Schischkin and Bobrov, 1960, 2002; Bobrov and Tzvelev, 2004). The complete flora consists of 22 000 pages and describes around 17 520 plant species belonging to 1676 genera and 160 families. (Shetler, 1967:136-166). N. I. Vavilov did not share the botanical view and in particular Komarovs monotypic species concept (Shetler, 1967). Vavilov was a spokesman for the geographic races

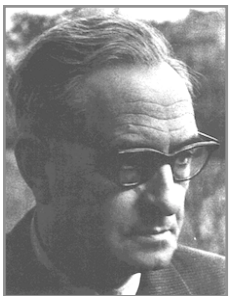
and ecotypes as a part of the species concept (Shetler, 1967:154). Komarov would tend to recognize *"local, morphologically homogenous races as distinct species"* (Shetler, 1967:150). With a few exceptions the Komarov flora is thus mostly without infraspecific taxa.

5.3.3 Bailey's flora of Cultivated Plants

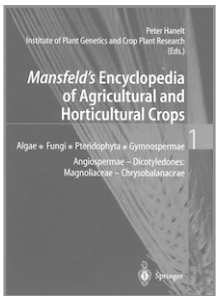
Liberty Hyde Bailey (1858-1954) contributed early to the description and naming of cultivated plants. In 1918 he started to use the term *cultigen*, and gave a formal definition in his *'Manual of Cultivated Plants'*: *"Plant or group known only in cultivation; presumably originating under domestication; contrast with indigen"* (Bailey, 1924, Glossary section). This book and others by Bailey contributed substantially to the taxonomy of cultivated plants (Banks, 1994).

5.3.4 Mansfeld's list of cultivated plants

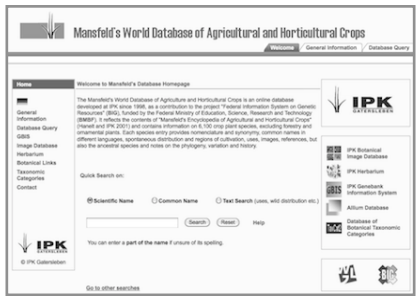
Rudolf Mansfeld (1901-1960) developed the first edition of the *'Preliminary enumeration of cultivated plant species'*. This list of names is today one of the most comprehensive references for genebanks (the other major reference being GRIN Taxonomy). The first edition of *"Vorläufiges Verzeichnis landwirtschaftlich oder gärtnerisch kultivierter Pflanzenarten (mit Ausschluss von Zierpflanzen)"* was published in 1959 as supplement (*'Beiheft'*) to the scientific journal *'Kulturpflanze'* (Mansfeld, 1959). An updated second edition was published in 1986 (Schultze-Motel, 1986). The first English edition (3rd edition) of the *"Mansfeld's Encyclopedia of Agricultural and Horticultural Crops (Except Ornamentals)"* was published in 2001 (Hanelt and IPK, 2001). This latest edition now covers 6121 cultivated plant species (Pistrick, 2003) and is available online at http://mansfeld.ipk-gatersleben.de/mansfeld (verified 25 Jan 2011) (Knüpffer et al., 2003a).



Rudolf Mansfeld (1901-1960)



Mansfeld's encyclopedia of agricultural and horticultural crops (Hanelt and IPK, 2001)



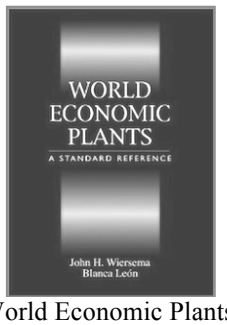
Mansfeld's World Database of Agricultural and Horticultural Crops (Knüpffer et al., 2003)

"Nowhere can you find more concise information on accepted scientific names, synonyms, common names, natural distribution, cultivation area, uses, history of cultivation and references of all agricultural and horticultural crops (except for ornamentals) in one place" (Pistrick, 2003:21).

5.3.5 GRIN Taxonomy for Plants

The USDA, ARS, National Plant Germplasm System (NPGS), Germplasm Resources Information Network (GRIN) provides an online searchable database on plant genetic resources maintained by genebanks in the USA. This database includes a section known as *'GRIN Taxonomy'* (http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl, verified 25 Jan 2011) including names for many economically important species conserved by US genebanks (and more). This list of names has become the *de facto* standard used by many genebanks worldwide for the naming of species (Wiersema, 1999; Terrell, 1977, 1986). The online GRIN Taxonomy includes now more than 50 000 (accepted) scientific names for more than 14 000 (accepted) genera. The *ISTA List of Stabilized Plant Names* (http://www.ars-grin.gov/~sbmljw/istaintrod.html, verified 25 Jan 2011) is largely based on the USDA GRIN Taxonomy.

Edward E. Terrell (1923-)

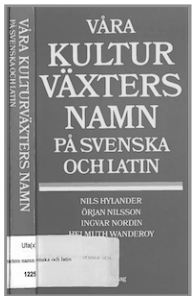


World Economic Plants (Wiersema & León, 1999)

GRIN Taxonomy, http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl

5.3.6 Swedish cultivated plants database (SCUD)

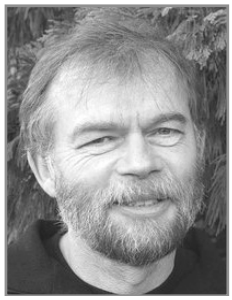
In 1948, Nils Hylander (1904-1970) compiled a list with botanical names for all ornamental plants cultivated in Sweden (Hylander, 1948). This book was later updated in a second edition in 1960 and a third edition in 1977 completed after his death. The third edition included some horticultural plants. In recent years, Björn Aldén maintains a list with botanical names of species cultivated in Sweden (Aldén, 1998; Aldén and Ryman, 2009). A searchable database version was provided online in 2005 (Aldén et al., 2005-2010), but lack of funding caused this online version to be closed down at the end of 2010 (Weibull, 2010).



Kulturväxters Namn på Svenska och Latin (Hylander et al., 1948)

Kulturväxtlexicon (Aldén, 1998)

Våra kulturväxters namn, ursprung och användning (Aldén and Ryman, 2009)



Björn Aldén

5.4 Georeferencing of genebank samples

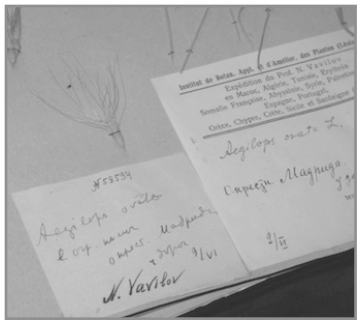
Accurate and complete documentation on genebank material is paramount for most operations involving the conservation or use of these genetic resources. In particular passport data regarding the georeferenced source locations where the original plant material was collected is important. A substantial part of the accessions in the genebank collections were collected before the hand-held geographic positioning system devices (GPS) became common. Even though the early collecting forms generally included instructions to record the coordinates, these coordinates are unfortunately often absent. From a total of 19 Nordic barley landraces included in the study reported with PAPER II, only four accessions had coordinates available from the Nordic SESTO genebank information system (NGB27, NGB775, NGB776, and NGB792). Georeferencing using various geographic information systems for the remaining 15 accessions resulted in the successful addition of coordinates for another 10 accessions based on the name and description of the source collecting locations. The georeferencing approach followed was described in PAPER II.

To georeference genebank accessions with missing geographic coordinates the first approach is running the name of the collecting location against a so-called gazetteer. A gazetteer is a list of place names and their coordinates. *Google Maps* (http://maps.google.org, verified 25 Jan 2011), *Getty Thesaurus of Geographic Names* (http://www.getty.edu/research/tools/vocabularies/tgn/,

verified 25 Jan 2011), and *BioGeomancer* (http://www.biogeomancer.org/, verified 25 Jan 2011) provide only a few examples of available online resources with gazetteer functions. The BioGeomancer (Guralnick et al., 2006) includes additional features to improve automatic georeferencing routines. Using the online application programming interface (API), skilled programmers can build an automatic routine to pass not only the name of the collecting site, but also written descriptions including statements such as *'10 km north-east of Alnarp, Skane, Sweden'* to calculate the georeferenced coordinates. The BioGeomancer prototype workbench (http://bg.berkeley.edu/latest/, verified 25 Jan 2011) can be used to georeference place name descriptions. Desktop GIS software such as DIVA-GIS (http://www.diva-gis.org/, verified 25 Jan 2011), Quantum GIS (http://www.qgis.org/, verified 25 Jan 2011), and GRASS (http://grass.osgeo.org/wiki/Georeferencing, verified 25 Jan 2011) include routines and plug-ins to perform efficient semi-automatic georeferencing of a large number genebank accessions in one single batch operation (Hijmans et al., 2001).

From the Vavilov Memorial at VIR, showing the collecting missions by Vavilov. Photo by Dag Endresen, April 2010.

Herbarium voucher for one of the VIR accessions, prepared by N.I. Vavilov. Photo by Dag Endresen in April 2010.



Close-up of herbarium voucher prepared by Vavilov. Photo by Dag Endresen, April 2010.

Another approach for georeferencing those collecting site descriptions difficult to match with the gazetteer approach could be to explore the germplasm accessions in the context of the respective collecting missions. If for example the description of the collecting site gives multiple hits using a gazetteer (or using BioGeomancer), some of the alternative locations can perhaps be eliminated through exploring the collecting locations for other accessions collected by the expedition on the same day. The whereabouts of the individual field worker on the respective collecting day for a difficult germplasm sample can for many uses provide a *'good-enough'* coordinate when reported together with a larger uncertainty estimate. Using the SESTO genebank information system, coordinate uncertainties are recorded as a circle with the radius reported in meters. This approach was tested at the Nordic Gene Bank (now Nordic Genetic Resource Center) with good success in 2002 (NGB, 2002a, 2002b; internal report, unpublished results). Meganck et al. (2006) described a similar approach developed as part the EU funded SYNTHESYS project (http://www.synthesys.info/, verified 25 Jan 2011).

A useful guide for georeferencing of biodiversity data is available from GBIF (Chapman and Wieczorek, 2006). Lessons can also be learned from Vavilov's descriptions of his own collecting expeditions with the book: *'Five continents'* (Vavilov, 1997), translated to English and published after Vavilov's death. In addition to the comprehensive overview of collecting sites visited by Vavilov, this book is written in an accessible style and provides a lively description of his collecting expeditions. Guarino et al. (1995) published somewhat more recent guidelines for conducting collecting expeditions for germplasm.

5.5 Species distribution models (SDM)

The ecological niche modeling approach for prediction of species distribution patterns (Stockwell, 2007; Franklin, 2010) has been used with good results for management of plant

genetic resources. One of the first ecogeographic studies of this type for crop germplasm was for wild peanut (species from the genus *Arachis*). Jarvis et al. (2003) calibrated species distribution models based on existing known occurrences of wild peanut, and used these models to estimate the likelihood of locating these wild peanuts in unexplored areas. Their study included 2175 georeferenced occurrences of wild peanut and found a good predictive performance using the FloraMap software (Jones et al., 1997; Jones and Gladkov, 1999-2005; Jones et al., 2002) and ecogeographic climate layers of 10-arc-minute resolution (approximately 18x18 km). This first study by Jarvis et al. (2003) to predict the whereabouts of wild peanut was a so-called desktop study. That is, the dataset with the 2175 georeferenced occurrences was split in a training set (1) used to calibrate the ecological niche model, and a test set (2) to validate the performance of the model.

Peanuts (*Arachis* sp.) by Dag Endresen, Jan 2011.

Capsicum sp. in Tunisia by Dag Endresen, Nov 2003.

Capsicum sp. in South Korea by Dag Endresen Oct 2010.

A second similar study by some of the same authors was designed to validate the ecological niche models for a rare species of hot peppers (*Capsicum flexuosum* Sendtn.) in the wild (Jarvis et al., 2005). This study was performed using the same software, FloraMap, and with the similar ecogeographic climate layers as used for the previous study. The chosen pepper species, *Capsicum flexuosum*, is one of the lesser-known species in the *Capsicum* genepool (Eshbaugh, 1993). Occurrence data from herbaria collections (in Paraguay, Argentina and USA) were combined with germplasm accessions from the USDA GRIN information system, resulting in a total of 19 unique populations identified and included in the study.

Accession Grif15020, collected 21 Mar 2002 in Canendiyu, Paraguay. Photo by Dr. Robert L. Jarret, USDA ARS in Georgia.

The collecting site for Grif 15020 is today surrounded by agricultural landscape (map from USDA GRIN, visited 5 Jan 2011)

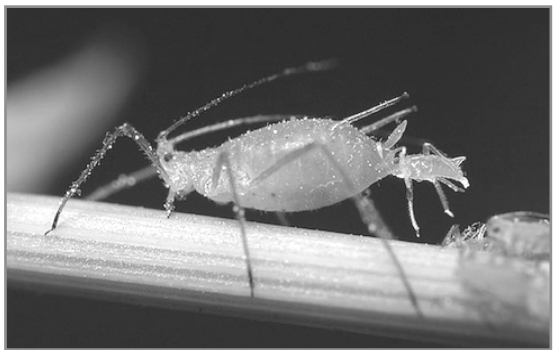
Capsicum flexuosum Sendtn. (Accession PI 631154, USDA GRIN) from prior collecting expedition, May 2001 in Paraguay.

To validate the predictions, 10 points from Paraguay were randomly chosen from the areas predicted by the model most likely to support populations of *C. flexuosum*, and 10 random points were chosen from areas with low predicted probability of finding *C. flexuosum*. These random points were restricted to locations within 4 km distance from an accessible road, and also to be previously unexplored. The seed collectors commissioned to explore each of these 20

locations (during March 2002), were not told whether they would visit a location with high or with low likelihood of finding the target species. Three of the selected locations proved to be inaccessible because of poor road quality (and 7 predicted absence points removed because the location itself was inaccessible to proper exploration). Six of the selected points were disturbed by human activities (so-called anthropogenic impacts, areas often transformed into soybean fields). From the remaining 12 points, the target species were found at 5 out of 7 predicted presence points (71% hit rate for predicted positives), and at 1 out of 5 predicted absence points (20% hit rate for predicted negatives, false negatives).

Based on these initial experiences with ecological niche modeling for germplasm from the crop genepools this approach was developed further to identify gaps in the conserved diversity for *ex situ* genebank collections. Scientists from the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) explored the sampled diversity of pearl millet (*Pennisetum glaucum* (L.) R. Br.) using FloraMap and identified potential major geographical gaps in the genebank collection (Upadhyaya et al., 2009). They identified many insufficiently sampled regions predicted by the models in FloraMap to be suitable for pearl millet to thrive. The authors suggest that these underexplored locations should be prioritized when planning for new collecting expeditions. Collecting germplasm from these locations was suggested to be the most efficient approach to achieve increased completeness of the ICRISAT pearl millet genebank collection. Ramírez-Villegas et al. (2010) published recently a similar gap analysis study for wild relatives of the *Phaseolus* genepool. The CIAT GIS department has established a team of experts on Gap Analysis. A total of 12 genepool analyses and 14 landrace analyses have been completed and published online (http://gisweb.ciat.cgiar.org/GapAnalysis/, visited 25 Jan 2011).

Sugar cane (*Saccharum*) in Venezuela by Flickr user: Rufino Uribe (CC-Attribution-Share-Alike).



Aphid by Wiki Commons user: "MedievalRich" (CC-Attribution-Share-Alike).

5.6 Species distribution models can predict the path for emerging pests

Ganeshaiah et al. (2003) explored species distribution modeling using the GARP (Anderson et al., 2003) and DIVA-GIS (Hijmans et al., 2001) software to develop a forecast system for emerging pests and crop diseases in India. The GARP software used a genetic algorithm for calibration of the ecological niche model, while the ecological niche-modeling approach they used with DIVA-GIS in this experiment was the BIOCLIM algorithm (Busby, 1991). The predicted distribution of the sugarcane (*Saccharum officinarum* L.) woolly aphid (*Ceratovacuna lanigera* Zehntner) insect pest was reported to be very similar, and the authors concluded that this approach to predict risk areas for new emerging pests and diseases showed great potential. The woolly aphid first appeared in the border area of the Indian provinces Maharashtra and Karnataka (Chakravarthy, 2004) in 2002. The emergence of this pest had a dramatic impact on sugar cane production (Joshi and Viraktamath, 2004). Ganeshaiah et al. (2003) predicted that the emerging woolly aphid pest was likely to spread to the nearby northern and southern dry zones, because the model indicated that the transitional belt connecting these two dry regions would be conductive for the pest. The likelihood of the pest to spread to the sugar cane growing areas to

the east of the first appearances, as well as to the coastal sugar cane growing areas, was concluded to be lower. The Mandya district was predicted to have the highest risk of receiving the pest, which should however, not proceed further south to Kerala and Tamil Nadu (Ganeshaiah et al. (2003). Joshi and Viraktamath (2004) confirmed the spread of the pest to the Mandya district (in agreement with the predictions), but also that the pest was observed further south in Kerala and Tamil Nadu (not in agreement with the predictions).

Kumarasinghe and Basnayake (2009) described the sudden infection of woolly aphid in Sri Lanka after dispersal from the nearby Indian subcontinent in 2006. With reference to the study made by Ganeshaiah et al. (2003), the authors suggested a similar prediction to be made for Sri Lanka to provide a forecast of the further emergence of the pest and as a basis for implementing efficient preventive measures against it.

Herborg et al. (2007) presented a somewhat different example of this type of study related to the prediction of the potential invasion risk by an alien species. Their study did not include a crop plant, but lessons can be learned nevertheless. The Chinese mitten crab (*Eriocheir sinensis* H. Milne-Edwards) is native to eastern Asia, established in Europe and recently emerged in Northern America. Using the GARP ecological niche modeling approach, Herborg et al. (2007) developed one model based on the known occurrences for the Mitten crab in Asia and a different model based on the known occurrences in Europe. Both models predicted roughly the same areas of North America as suitable for the Mitten crab. A similar approach can perhaps be used to predict the risk areas for weeds and crop pests or pathogens in agricultural systems.

Chinese mitten crab (*Eriocheir Sinensis* H. Milne-Edwards) from the river Elbe, near Brandenburg in Germany.

Photo by Christian Fischer (CC-Attribution-Share-Alike).

5.7 The impact of climate change on plant genetic resources

Jarvis et al. (2008) explored the effect from global warming for the wild relatives of peanut (*Arachis*), potato (*Solanum*), and cowpea (*Vigna*). The estimated species distribution for the present climatic conditions was compared to the projected species distribution under a future climatic scenario for 2055. These climatic data layers are available from the online WorldClim information system (http://worldclim.org, verified 25 Jan 2011; Hijmans et al., 2005). The experiment was set up using the BIOCLIM algorithm (Busby, 1991) as implemented with DIVA-GIS (http://diva-gis.org, verified 25 Jan 2011). Only species with more than 10 georeferenced occurrences available were modeled (a total of 210 species). Under these experimental conditions an estimated 16-22% of the species were predicted for extinction, and most of the remaining species predicted to lose more than 50% of their distribution range. The authors suggested urgently starting the collecting of the species with the most severe predicted loss of distribution range.

"Recent studies indicate that increased frequency of heat stress, droughts and floods negatively affect crop yields and livestock beyond the impacts of mean climate change, creating the possibility for surprises, with impacts that are larger, and occurring earlier, than predicted using changes in mean variables alone. This is especially the case for subsistence sectors at low latitudes. Climate variability and change also modify the risks of fires, pest and pathogen outbreak, negatively affecting food, fiber and forestry" (IPCC, 2007).

5.8 Linking ecogeographic data with evaluation data

Another type of study (and the central topic of this thesis) involving ecogeographic data analysis is the exploration of the predictive link between the ecogeographic description of the original collecting site for germplasm accessions and their useful traits.

John P. Peeters and J. Trevor Williams suggested that the ecogeographic description of the collecting sites can be used for trait mining: *"In order to have some grounds for selection of material for evaluation purposes, the first and most important level of information in general relates to passport data. (...) By using passport data, samples can be taken from different putative 'genepools' on the basis of their different ecogeographic backgrounds. The number of samples selected for evaluation will depend on the levels of finance available"* (Peeters and Williams, 1984:29). As a basis for these suggestions by Peeters and Williams, the authors expressed acknowledgement to the contributions from Otto Frankel. Peeters et al. (1990) made a study to test this hypothesis for the selection of samples of wild barley (*Hordeum spontaneum* K. Koch) and the assumed predictive link between tolerance to salinity and rainfall pattern. The method they used was simple regression between these two variables, and their results concluded that: *"although ecological variables in the site of origin can be useful in predicting genetic characteristics in the samples, the use of such data is neither simple nor precise"* (Peeters et al., 1990:110). They further found that regression between these two variables was much more successful when performed for subsets stratified by country of origin than for the larger combined dataset.

Robert J. Hijmans et al. (2003) made a similar conclusion when they tried to find a predictive link between temperature data from the source collecting sites and the frost tolerance for a total of 1646 wild potato (*Solanum* spp.) samples from 87 species and 12 countries in the Americas. Their data modeling approach was that of GLM regression. The link between the ecogeographic variable and the trait variable was concluded to be weak and complex (Hijmans et al., 2003).

Shelly H. Jansky, Reinhard Simon and David M. Spooner conducted a series of trait-mining experiments with evaluation data on pests and diseases for wild potato (*Solanum* spp). With the first experiment they explored the link between ecogeography and resistance to white mold (*Sclerotinia sclerotiorum* (Lib.) de Bary), only to find a weak correlation between the trait and the climate data. This was one of the first studies to explore the predictivity of ecogeographic data for disease resistance in germplasm material (Jansky et al., 2006). The second experiment explored the link between ecogeography and early blight (*Alternaria solani* Ellis and G. Martin), where they found a significant predictive correlation with both the precipitation and the altitude pattern of the collecting site (Jansky et al., 2008). In a third study they found only a weak predictive link between ecogeography and resistance to the Colorado potato beetle (*Leptino decemlineata* Say) for wild potato germplasm samples (Jansky et al., 2009). The last of these fours experiments was a much larger study to explore the predictive link for a total of 32 crop pests and diseases in wild potato. Using two different classifiers, Support Vector Machines (Cortes and Vapnik, 1995) and Random Forest (Breiman, 2001), they found a predictive relation for six of these pest and disease traits (Spooner et al., 2009).

5.9 Agroecological classifications by Nikolai I. Vavilov (1887-1943)

Vavilov devoted significant efforts to the classification of the cultivated plants into agroecological groups for the purpose of structuring and identifying the raw materials for crop improvement (Vavilov, 1957, 1964; Dorofeev, 1992). He classified regions of the world into *"agroecological districts"* and explored the characteristic developments and adaptations of cultivated crops linked to these agroecological districts: *"List of main characters and properties considered in the elaboration of an agroecological classification adapted to annual cereals, grain leguminous crops and flax"* (Vavilov, 1957: 80). Knüpffer et al. (2003b) described Vavilov's approach to study and identify useful diversification based on ecogeographical adaptations: *"This classification provides an orientation or entry point for studying the world resources of barley in more detail in search of interesting traits, such as drought resistance, naked grains, or resistance to lodging"* (Knüpffer et al., 2003b:71). Most of the works by Vavilov on ecogeographic adaptation of useful traits are available only in Russian and are thus somewhat less accessible to Western scientists. It is expected that the further translation and improved accessibility of Vavilovs writings on this topic would lead to a deeper understanding of the principles behind recent developments including the Focused Identifications of Germplasm Strategy (FIGS). Likewise the crop-based ecogeographical surveys published in Russian as part of the Cultivated Flora of the USSR (Flora of Cultivated Plants) are expected to contribute to a more widespread understanding of this topic - if they are translated to English. An English translation of vol. 1, the Wheat Flora (Dorofeev et al., 1979) is underway (*cf.* Morrison et al., 2000; Knüpffer et al., 2002). Today most of these important reference works remain readily accessible only to scientists who master the Russian language.

The work by Vavilov on the agroecology of cultivated plants provided some of the early concepts behind the FIGS approach including the assumption of a link between the traits of cultivated plants and the environmental conditions where they where developed and cultivated (Vavilov, 1957). However it is not unlikely that much of the work made by Vavilov is now remade in present times because of the lack of easy access for western scientists to most of the scientific publications written by Vavilov on this topic. Initiatives to provide translations from Russian to English (Morrison et al., 2000; Knüpffer et al., 2002) suffer lack of funding for the publication costs.

5.10 Conclusion

Ecogeographic classification and studies on genebank material provides essential information for rational conservation and use of these plant genetic resources. Such ecogeographic studies are required before a trait mining experiment can be made. The georeferencing of genebank material to assign accurate geographic coordinates is the required first step to enable the extraction of climate data for the genebank accessions from ecogeographic datasets. The lack of geographic coordinates for germplasm samples provides an equal requirement for ecological niche modeling experiments including the so-called 'gap analysis'. It is recommended that ecogeographic studies of genebank material and herbaria samples, including the taxonomic classification and georeferencing be given higher priority.

CHAPTER 6

6. Core collections and FIGS

During the second half of the 1970s and the 1980s a large number of accessions representing a vast amount of plant genetic diversity were collected and deposited into the new genebank facilities established during this time. However many of these genebank collections suffer gaps in the descriptive information required to guide the rational and efficient use of this germplasm in crop improvement. Paradoxically another major obstacle to the exploitation of the new genebank collections was the large number of accessions. Systematic field trials to gather characterization and evaluation data were initiated, but the large number of accessions for a specific crop was far larger than the capacity of the individual screening projects. Jack R. Harlan proposed already in 1971 to rationalize the task of germplasm evaluation by deriving subsets that he called *'active working collection'* and to focus the screening efforts on the accessions in these subsets (Harlan, 1972 *cf*. Spagnoletti Zeuli and Qualset, 1993). Marshall and Brown (1975) proposed a sampling strategy to derive a subset with the maximum diversity in a limited number of samples. Frankel (1984) proposed that this subset should be named *'Core collection'*. The Core collection concept attracted substantial interest in the genebank community. Mackay (1986:59, 1990, 1995) proposed that the Core collection should be designed in such a way as to select targeted accessions for the specific trait property requested by users such as plant breeders. This sampling strategy has a different goal than the core collection and was later to be known as Focused Identification of Germplasm Strategy (FIGS; Mackay and Street, 2004).

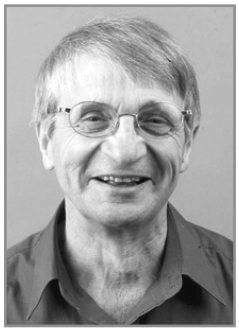
6.1 Core collections

The challenges caused by large size of genebank collections and lack of descriptive data on genetic resource collections were some of the reasons for the introduction of the core collection concept by Sir Otto Herzberg Frankel (1900-1998) (Frankel, 1984; Frankel and Brown, 1984a, 1984b). When Frankel prepared his original paper on the *Core collection* concept, he was familiar with the work by Donald Robert Marshall and Anthony (Tony) Hugh Dean Brown (1941-) on germplasm sampling strategies (Marshall and Brown, 1975) and called upon Brown to read the draft manuscript where he introduced the core collection approach (Palmer and Doyle, 2009). Some years earlier Marshall and Brown presented their proposed sampling approach at the 1973 FAO/IBP Technical Conference on Crop Genetic Resources: *"Our purpose here is to formulate, as far as possible with current information, a quantitative sampling theory for genetic conservation which permits the collection of the maximum amount of genetically useful variability in the target species while keeping the number of samples within the practical limits discussed above. However, before we can consider the problem of optimal strategies in detail we need to define (i) and appropriate measure of genetic diversity - the parameter we wish to maximize, and (ii) what we regard as genetically 'useful' variability"* (Marshall and Brown, 1975:55). Frankel and John (Jack) Gregory Hawkes (1915-2007) were the editors for a book based on the contributions from the 1973 FAO/IBP conference (Frankel and Hawkes, 1975). It is likely to assume that the manuscript by Marshall and Brown (1975) made an important contribution to the formulation of the concept for core collections by Frankel in 1984.

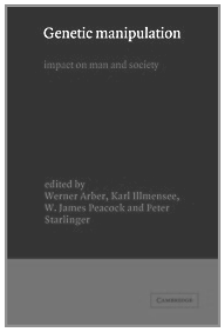
Frankel was 84 years in 1984, and a renowned and distinguished international scientist with an important role as the front-figure for the urgent need to collect genetic resources under the regime of genetic erosion. It is thus reassuring to our image of Frankel today to consider that he

deliberately intended the concept of the *reserve collection* for the purpose to hold all the samples not selected for the *core collection* (Frankel, 1984). Palmer and Doyle (2009:5) noted that Frankel often said to Brown: *"Tony, I have you to blame for the reserve collection"*, and that Frankel felt that the core collection concept did not make as strong call to genebank managers, for pruning their collections, as he originally intended. We can thus deduce that Frankel was of the opinion that some of the genebank collections would need some rationalization and reduction of redundant samples.

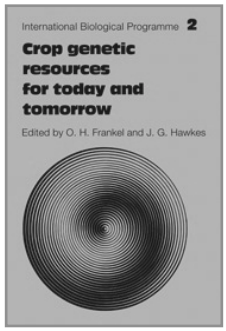
Sir Otto Herzberg Frankel (1900-1998)



Anthony (Tony) Hugh Dean Brown (1941-)



Genetic manipulation, impact on man and society (Arber et al., 1984)



Crop genetic resources for today and tomorrow (Frankel & Hawkes, 1975)

6.2 Sampling strategies

In 1984 Frankel coined the term and defined concept of the *Core collection*, set out to maximize the diversity in a limited subset of the entire germplasm collection to get a more manageable subset (Frankel, 1984). A core collection seeks to represent most of the genetic variation present in the original collection in a *'core'* subset of 5-10% the size of the original. Core collection sampling strategies use statistical approaches to maximize measures of diversity using a variety of input data including collection site descriptors, agro-morphological traits and molecular marker data. The development of new core collection sampling strategies has been well covered in the scientific literature (examples include: Marshall and Brown, 1975; Brown, 1989a, 1989b; Peeters and Martinelli, 1989; Spagnoletti Zeuli, and Qualset, 1993; Charmet and Balfourier, 1995; Hodgkin et al., 1995; Hintum et al., 1995; Balfourier et al., 1998; Hintum, 1999; Hintum et al., 2000; Upadhyaya et al., 2001; Franco et al., 2005).

The proposed core sampling strategies include:

- Random sampling (selecting random samples from the complete dataset)
- Systematic by chronology (such as sorted by accessions number)
- Stratified sampling by geographic origin (country of origin) or other grouping
  - The same number of samples from each group (C strategy)
  - Proportional to the number of samples from each country or group (P strategy)
  - Proportional to the logarithm of total samples from each country/group (L strategy)
- Stratified by canonical/latent variables (clusters)

(Brown, 1989; Hintum et al., 2000:18)

Other stratified sampling strategies was based on molecular marker data (or characterization and evaluation data, C&E):

- Empirical stepwise model to maximize observed diversity in each step (M strategy)
  - By observed allelic diversity (molecular genetic data)
  - By observed phenotypic trait, qualitative/quantitative (C&E data)
- Maximize the number of different alleles in the core collection (H strategy)
  - The H strategy can also be extended to use C&E data

(Schoen and Brown, 1995; Hintum et al., 2000:19)

6.3 Bias towards representing diversity rather than usefulness

The core collection approach may lead to the exclusion of rare alleles from the core subset derived from crop collections (Brown, 1995; Mackay, 1995). The purpose of the core subset selection is to sample a collection of germplasm to get a reduced set with minimum similarity between the entries to maximize the genetic diversity in the smaller set. Brown (1989b) calculated that a subset of 10% would include an acceptable proportion of the genetic diversity (under the assumption of the neutral allele model (Ewens, 1972). The smaller set is designed to represent the larger germplasm collection. The core collection strategy is thus not designed or expected to improve the chance of finding a particular targeted trait property, including for example the resistance to a crop pathogen. The core subset represents a bias towards diversity rather than usefulness (Brown, 1995).

6.4 Focused Identification of Germplasm Strategy (FIGS)

The FIGS is an approach used to select subsets of germplasm from genetic resource collections in such a way as to maximize the likelihood of capturing a specific trait at a higher frequency than if the subset was selected at random. Michael Mackay proposed the first version of this subset sampling strategy as a core collection sampling strategy (Mackay, 1986, 1990). The current naming as *'Focused Identification of Germplasm Strategy'* (FIGS) was proposed some years later when Michael Mackay and Kenneth Street developed the strategy further (Mackay and Street, 2004). The FIGS subset is designed to meet the needs of a plant breeder approaching a genebank for sources of genetic diversity for crop improvement programs in a specific trait. *"The characters for which the breeders turn to gene banks often are very rare variants or combinations, partly because the characters that are common in a crop species are likely to be in a breeder's working collection already"* (Duvick, 1984 *cf*. Brown, 1995:8).

The genebank collections of plant genetic resources are today a valuable source of new genetic variation for economically important traits, including resistance to crop diseases. New sources of useful crop traits are often identified through evaluation in field trials, some of which may require special treatments such as inoculation for screening purposes. This has tremendous implications as it incurs costs. Further, the number of relevant accessions in genebank collections available to be evaluated for a specific trait is often substantially larger than the capacity or resources of the evaluation project. Thus, finding the genebank accessions that are most likely to posses the desired trait can be compared to searching for a needle in a haystack.

6.5 The early origins of the FIGS concept was presented in the 1980s

In 1986 at the Fifth Assembly of the Wheat Breeding Society of Australia Michael C. Mackay presented his first experiences with the sampling approach later to be known as *'FIGS'*: *"Users are also encouraged to consider the 'predictive' value of existing descriptors before making requests. A combination of origin, growth type, ear emergence and plant height would fairly accurately predict the type of environment and accession came from. This knowledge could then be used to select targeted groups of accessions from environments likely to have attributes being sought"* (Mackay, 1986:59).

In 1990 Michael Mackay presented what is now considered as the founding paper for the FIGS concept. This paper was presented at a scientific symposium organized at ICARDA (Aleppo, Syria), and the contributions from this symposium were compiled and published in a book format (Srivastava and Damania, 1990). He does not use the FIGS term, but describes the concept as a *"strategic approach"* to derive a smaller subset of genebank accessions targeting a specific trait for the purpose of germplasm evaluation. This new approach was presented as a further development of the sampling strategy implemented by the Core collections: *"This*

approach is based on the 'core collection concept'" (Mackay, 1990:22). It was proposed that information about the geographic origin for accessions could be used to deduce *"the type of environment in which an accession evolved or was selected"*. The type of environments positively linked to a specific target trait character would then be used to select genebank accessions based on the type of environments at the locations of their geographic origin (Mackay, 1986, 1990). *"This technique has been used with considerable success to select material with tolerance to pre-harvest sprouting, resistance to cereal cyst nematode, and tolerance to boron toxicity (Mackay, 1986)"* (Mackay, 1990). The strength of this new sampling approach as compared to the Core collection approach was that the targeted subset was derived in collaboration with the plant breeder and targeting maximal diversity for the specific trait property the breeder was searching for, rather than maximal diversity across the entire genebank collection.

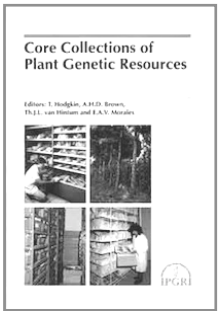
When Toby Hodgkin summarized the status of core collections at the EUCARPIA/IBPGR symposium on Crop Networks held at Wageningen in 1990, he described Mackays subset sampling approach as: *"The approach taken by Mackay (1986, 1988* [I believe that this reference should be '1990']*), in using the core concept for improved evaluation and utilisation of wheat in the Australian winter cereals collection, differs considerably from the other examples given. (...) This procedure is complemented by a specific attribute programme for which the genebank and potential user make predictive decisions about the origin of germplasm, which might possess the desired character in terms of soil type, climate, maturity, etc. This is used to identify sets of accessions which are likely to contain maximum variation for the character in question"* (Hodgkin, 1991:46).

6.6 One core or many?

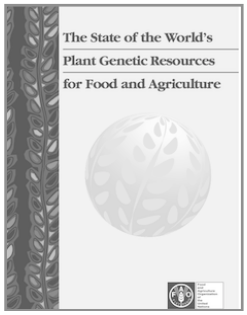
In August 1992, IBPGR, CGN and CENARGEN convened a workshop on *'Core Collections: Improving the Management and Use of Plant Germplasm Collections'*, hosted by *Centro Nacional de Pesquisas de Recursos Genéticos e Biotecnologia* (CENARGEN) in Brasilia, Brazil. The contributions to this workshop was first published as conference proceedings (IPGRI, 1992), and compiled as a hardcover reference book (Hodgkin et al., 1995) due to the interest this report generated. The core collection concept was at this time period, a central topic on the agenda of the plant genetic resources community, as can be illustrated by the coverage in the Report on the State of the World's Plant Genetic Resources (SOTW; FAO, 1997).

Michael Mackay

One Core or Many? (Mackay, 1995)



Core Collections of PGR (Hodgkin et al., 1995)



State of the World's PGRFA (FAO, 1997)

Michael Mackay contributed to this workshop, and to the following book, with a manuscript titled *'One core collection or many?'* (Mackay, 1995). Here Mackay questioned the widely recognized assumption at the time, that evaluation data is required for rational use of genebank material by plant breeders. This assumption generated general efforts to start the systematic evaluation of core collections, or even complete genebank collections. He proposed that passport data (and other types of accession level information) can also be used to assist the

selection of target accessions for a specific trait; and with significantly less costs than the systematic evaluation of core collections for trait properties that we sometimes don't even know in advance if the breeders will ask for. *"Breeders require rapid identification of desirable attributes and immediate access to seed samples"* (Mackay, 1990:200). *"[G]ermplasm users often request a set of accessions that is likely to contain a characteristic not previously described"* (Mackay, 1995:203). Mackay proposed to use a terminology including the word *'core'* like for example *"specific purpose (attribute) core sets"* for describing target set subsets derived by his proposed approach (Mackay, 1995). However Brown advised with another manuscript in the same book against extending the usage of the established *'core collection'* term in this way. Instead Brown suggested using terms such as *"acid-soil tolerant set"* (*'[special purpose] set'*) (Brown, 1995). This contribution by Mackay in 1995 provides the early theoretical foundation for the sampling method later to be known as FIGS.

6.7 Project funding from GRDC

In 2002 a four-year grant was approved from the Grains Research and Development Corporation (GRDC) located in Australia, to develop the proposed sampling strategy further (the online GRDC Investment Portfolio and Annual Reports are included in the reference list to this chapter). The GRDC funded project *"Technologies for the targeted exploitation of the N I Vavilov Institute of Plant Industry (VIR), ICARDA and Australian bread wheat landrace germplasm for the benefit of the wheat breeding programs of the partners"* was established at ICARDA in collaboration with the Vavilov Institute in St Petersburg, and the Australian Winter Cereal Collection (AWCC) in Tamworth. The success of this project has ensured continued funding from GRDC during the recent years.

6.8 Focused Identification of Germplasm Strategy (FIGS)

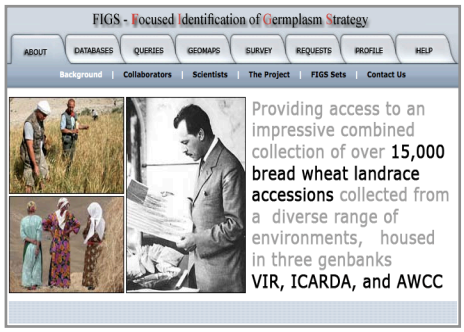
The term *'Focused Identification of Germplasm Strategy'* was already formulated before 2003, probably as early as around 1995. Around the years 2003 to 2004 Kenneth Street (ICARDA) and Michael Mackay developed a project concept note under the title: *"Revolutionizing Plant Genetic Resources Management, Trait Discovery & Utilization"* (Street and Mackay, 2003-2004). This 4-page document describes a project proposal presented by Kenneth Street (ICARDA), Michael Mackay (AWCC), Jan Konopka (ICARDA), Bent Skovmand (NGB), and Olga Mitrofanova (VIR). During the second half of 2004 the FIGS project was presented at 4 different international crop conferences. In July FIGS was presented to the 7th International Oat Conference hosted by MTT in Jokioinen, Finland: *"The most common strategy, the core collection strategy, is one designed to catch what you can, rather than catch what you need. Here we propose a new strategy that more fully exploits existing information to maximize utilization of genetic resources"* (Skovmand et al., 2004:35). In September, FIGS was presented at the 11th Wheat Breeders Assembly, organized in Canberra, Australia (Mackay and Street, 2004). Another conference paper on FIGS was presented at the ASA CSSSA SSSA 2004 International Meeting held in Seattle, Washington State, USA (Street et al., 2004).

6.9 FIGS project web site

In 2005 a dedicated web portal was developed for the FIGS project (Konopka et al., 2005-2007; http://www.figstraitmine.org/, verified 25 Jan 2011). One of the purposes with this portal was to construct and present the first FIGS subsets using the targeted sampling approach. The first FIGS sets to be included were a Drought set, Salinity set, Powdery Mildew set, Russian Wheat Aphid set. Each of these sets was made available from an online search interface integrated into the portal.

Michael C. Mackay

Kenneth A. Street



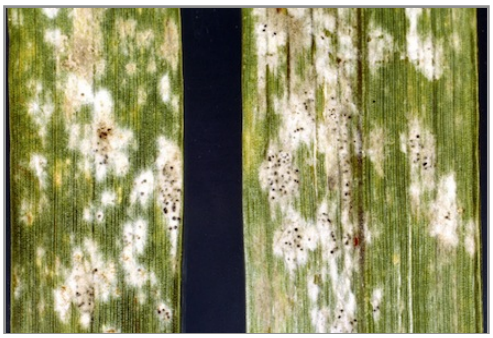
Bread Wheat Landrace Database (FIGS data portal), available at www.figstraitmine.org

6.10 Allele mining

The application of FIGS for allele mining is a similar approach as trait mining for the identification of genebank accessions for a target property. Trait mining use the FIGS approach to predict observable phenotypic (morphological) or agronomical trait characters. The allele mining approach is a further development based on FIGS, aimed at linking data from molecular biology studies such as molecular marker data with ecogeographic datasets (Bhullar et al., 2009). The allele mining approach thus use molecular methods to identify samples with a higher probability of holding useful properties. Allele mining is also known as gene mining (Vincente et al., 2006).

6.10.1 Allele mining at the Vavilov Institute

The wheat department at the Vavilov Institute was a partner of the GRDC funded FIGS project. The primary focus from the team of Russian scientists has been on allele mining (Strelchenko et al., 2003; Strelchenko et al., 2004; Balfourier et al., 2007; Strelchenko et al., 2008; Mitrofanova et al., 2009). *"Over thousands of years landraces of hexaploid wheats, with a genome composition of AABBDD, have developed under a variety of different edaphic and climatic environments. This has led to the evolution of a large number of ecotypes adapted to specific local environments. In the past attempts have been made to describe the eco-geographical differentiation of wheat using morphological and agronomical traits (Palmova 1935; Vavilov, 1964). Recent developments using PCR based methods have allowed fast and effective approaches for examining plant polymorphism at the DNA level"* (Strelchenko et al., 2004).



Powdery mildew (*Blumeria graminis* (DC.) Spear) on barley. Photo made available by Clemson University (CC-Attribution-3.0).
Source: forestryimages.org/browse/detail.cfm?imgnum=1436028 & imgnum=1233233

Mildew on bread wheat, NGB 6676 (Photo by Axel Diederichsen, NordGen)

6.10.2 Allele mining at the University of Zurich (Powdery mildew, *Pm3*)

The extended application of the FIGS approach for allele mining was further developed at the Institute of Plant Biology, University of Zurich by Navreet Kaur Bhullar at the research group coordinated by Beat Keller (Kaur et al., 2008a, 2008b; Bhullar et al., 2009, 2010a, 2010b; Street et al., 2008). The primary objective was to identify desired sequence diversity in wheat

landraces for sources of resistance to powdery mildew (*Blumeria graminis* (DC.) Spear f.sp. *tritici*) using FIGS (Mackay and Street, 2004) together with Allele mining (Latha et al., 2004; Berger, 2004; de Vicente et al, 2006). The focus of their work was the *Pm3* locus in wheat (*Triticum* spp.). The allele mining approach was formulated as a strategy to identify new alleles at a known locus. The identification of novel resistance alleles at resistance loci in the genepool of cultivated crops provides an important element of crop improvement. Allele mining with FIGS was demonstrated to be a very successful and efficient approach to identify such novel sources of resistance in genebank collections (Kaur, 2008; Bhullar, 2009).

6.11 Some examples of core collections relevant to targeted sampling

Two examples for the design of core collections for a target trait are presented with 6.11.1 and 6.11.2. The barley core collection was designed as a virtual core collection including genebank accessions of barley from many genebank collections. This concept provides a useful example for trait mining experiments to build a combined dataset with material from many genebanks such as is provided by the GeneSys (http://www.genesys-pgr.org/, verified 2 Feb, 2011) and by the progress on interoperability between genebank datasets (ABSTRACT 2-5).

6.11.1 Mini core strategy to develop a target subset

Holbrook and Dong (2005) proposed a way of using the core selection strategy to increase the hit rate for the identification of useful alleles for traits that are particular difficult or expensive to measure. Upadhyaya and Ortiz (2001) had previously shown that a so-called mini core subset containing only 1% of the total accessions in the entire collection still might capture most of the useful target variation. Recall that Brown (1989b) suggested that the core subset should include 10% of the accessions in the entire set. What Holbrook and Dong (2005) did was to split the accessions of the US peanut (*Arachis hypogaea* L.) core collection (831 accessions) into sub-groups using a standard cluster analysis method (*'Ward's minimum variance cluster analysis'*, Ward, 1963) using 16 morphological characters (8 traits measured respectively before and after harvest). The mini core was then derived as a random selection of 10% accessions from each of these sub-groups. The theory was that the results from a disease screening in the mini core (with 1% of the total accessions, 112 samples) would guide the experimenter to identify the target trait in the core subset (10% of the total accessions, 831 samples). The results from the trait screening of the mini-core were used to determine which of the sub-groups in the core set to be examined during a second screening experiment. In this study the entire core set was screened, however a simulation of the two-stage screening approach was used. The trait dataset tested for leaf spot (*Cercospora arachidicola* Hori and *Cercosporidium personatum* Berk. & M. A. Curtis) indicated that the screening of only 54% of the core collection would have identified 90% of the leaf-spot resistant accessions in the core collection. Similar results were found in the same study for tomato spotted wilt virus (TSWV, genus *Tospovirus*), but not for resistance to the peanut root-knot nematode (*Meloidogyne arenaria* Neal) and the aflatoxin contamination caused by fungal infection (*Aspergillus flavus* Link ex Fries (NRRL 3357) and *Aspergillus parasiticus* Speare) (Holbrook and Dong, 2005).

6.11.2 Thematic core collections for improved capture of rare alleles

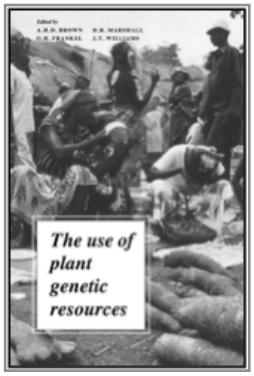
Some concerns are raised that the ever-increasing number and size of genebank collections around the world are starting to cause even the core collections to include more samples than the genebank customers are willing to receive. Pessoa-Filho et al. (2010) has proposed to create thematic core collections for each target trait and other specific purposes. They proposed to use molecular markers and pairwise genetic distances to design thematic core collections in such a way as to maximize the diversity for the target trait. This approach is also a two-stage approach using prior trait measurements for the target trait to maximize the diversity in the core subset for the trait in question. The purpose of this approach is to derive smaller core collections while

maintaining the same level of diversity for a specific trait. Mackay has previously proposed a similar strategy with the manuscript bearing the title *"One core or many"* (Mackay, 1995).

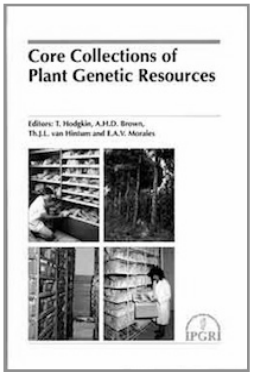
6.11.3 The barley core collection is a synthetic core collection

The development of crop core collections from an individual genebank collection does not help the genebank customer interested in a representative subset from any genebank collection. The genebank user will still need to request accessions from multiple genebanks with different implementations of the core collection strategy (or no implementation of this subset approach at all). The barley core collection is designed to guide the genebank customer to a relevant smaller subset of barley accessions from all partner genebanks sharing information on their barley accessions with the central database hosted by IPK Gatersleben in Germany (Hintum et al., 1990; Knüpffer and Hintum, 2003).

The number of books published with a focus on Core collections and their sampling strategies witnesses the great interest in this concept:



The use of plant genetic resources (Brown et al., 1989)



Core Collections of Plant Genetic Resources (Hodgkin et al., 1995)

Core collections for today and tomorrow (Johnson and Hodgkin, 1999)

Core Collections of Plant Genetic Resources (van Hintum et al., 2000)

6.12 Software implementations for the sampling of core collections

The number of dedicated software tools developed to assist the core subset selection provides a similar proof of the great interest in core collections.

6.12.1 MSTRAT (Gouesnard *et al.*, 2001)

The MSTRAT software is available for Windows and Unix-like systems (as the Linux and Mac OsX). The graphical user interface (GUI) is based on R and Tcl. There is also a command line alternative. MSTRAT is distributed as C source code (open source, creative commons) and was developed at INRA in France. The MSTRAT toolkit takes plain space separated ASCII text files as input (and output). It is based on the so-called *M-strategy* to maximize the number of observed alleles at the marker loci as introduced by Brown (1989b) and further described by Schoen and Brown (1993). The MSTRAT software has implemented both an extended version of the M-strategy, including as input qualitative and quantitative trait data, as well as the original M-strategy input only from allelic diversity at a specific loci.

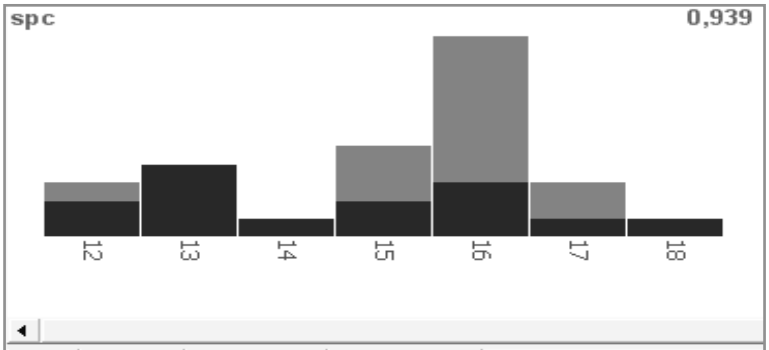
6.12.2 PowerMarker (Liu and Muse, 2005)

This software is available for download from: http://statgen.ncsu.edu/powermarker/ hosted by the North Carolina State University (USA). This software was not originally designated to make a core collection subset, but includes a number of routines to analyze genetic marker data. PowerMarker requires a Windows system and is developed with Visual C# and Visual C++, and runs in the Microsoft .NET framework. The current version is 3.25 and was released on 5 February 2006.

6.12.3 PowerCore (Kim *et al*, 2007)

PowerCore is a software implementation designated to make a core selection following the M strategy to maximize the allele richness. PowerCore is developed by the South Korean genebank with Microsoft Visual Studio as C# code and runs in the .NET framework. The software can be downloaded for free as a Windows system compatible binary from: http://www.genebank.go.kr/eng/PowerCore/powercore.jsp.

PowerCore requires a Windows system to run. Work is in progress to port the entire .NET framework from Microsoft to Linux and Mac, however no work is in progress to port the PowerCore software independently to the Linux and Mac environment (Figure 6.1).



	Class	Core Count	Entire Count
▶	12	2	3
	13	4	4
	14	1	1
	15	2	5
	16	3	11
	17	1	3
	18	1	1

Figure 6.1: Screenshot from the PowerCore software interface (using the same dataset with wheat landraces as analyzed in PAPER III). The image illustrates accessions split for one of the input trait variables (spc) were the lower dark blue sector of the bars indicate the sampled accessions.

6.12.4 Core Hunter (Thachuk *et al*, 2009)

Core Hunter is a recent software tool designed to develop a core subset developed with Java and free to download and use. Core Hunter can be downloaded from: http://www.corehunter.org/ (verified 14 Jan 2011), and from: https://cropforge.org/projects/corehunter/ (verified 14 Jan 2011).

6.13 Software implementations are under development for FIGS

Appendix 4 provides the algorithms and MATLAB code used in PAPER III and IV for the sampling of targeted genebank samples using FIGS. These MATLAB scripts can be seen as a first step towards a software implementation for trait mining with FIGS. An initiative is made for migrating of these algorithms to the R software platform. MATLAB is proprietary software with a high economic cost that will cause a problem for many genebanks and in particular those based in developing countries. Implementation of FIGS with the open source R platform distributed free of license costs will thus provide more wide access to this new software tools.

6.14 Conclusion

The sampling strategies developed for the Core collection concept (Frankel, 1984) will still have great utility for genebank managers with the task to maintain genetic diversity. Trait mining using the FIGS approach was proposed as a strategy to identify specific trait properties when these are requested by plant breeders (Mackay and Street, 2004). A bibliography with an overview of published studies using the FIGS approach is included as appendix 3.

CHAPTER 7

7. Trait mining

The FIGS strategy uses different methods to link a target crop trait with the ecogeographic parameters of the original collection site. The work presented as PAPER II, III, and IV provides some of the first experimental proofs for the FIGS concept, and also the first application of multi-way data analysis methods both for trait mining using FIGS and for analysis of genebank datasets in general.

7.1 Trait mining experiments at ICARDA for Sunn pest and RWA

ICARDA was the coordinator of the GRDC funded project (started around 2002) to develop the FIGS concept further. The results from the first experiments using FIGS were during the recent years completed and prepared for scientific publication. A database was compiled including 16 000 accessions of bread wheat (*Triticum aestivum* L.) and durum wheat (*Triticum turgidum* L.) from ICARDA, Vavilov Institute and AWCC. This database was used to derive targeted subsets for Sunn pest and Russian Wheat Aphid using the FIGS approach as described in more detail below (7.1.1 and 7.1.2) (Street et al., 2008). These two studies provided the first experimental evidence in support of the FIGS sampling strategy. The modeling approach was primarily a step-wise approach with decision rules based on expert knowledge and experience.

Sunn pest feeding on wheat. Photo made available by University of Vermont.

Russian wheat aphid feeding on wheat. Photo by Frank Peairs, Colorado State Uni CC-Attrib-3.

Russian wheat aphid. Photo by Frank Peairs, Colorado State Univ, CC-Attrib-3.

Aphid colony. Photo by John A. Weidhass, Virginia State Univ, CC-Attrib-3.

7.1.1 Sources of wheat resistance to Sunn pest using FIGS

The FIGS set for Sunn pest (*Eurygaster intericeps* Puton) was derived in a step-wise and rule-based manner. The first step was to limit the bread and durum wheat samples to accessions originating from areas where Sunn pest has been reported in the past (1). This rule made a reduction of the full set down to less than half of the samples from the full set. The second step sampled randomly one single accession from each collecting site (2). Very dry climate (3) and harsh winters (4) are both reported to limit the risk of Sunn pest crop damages and used as the third and forth filter. The number of samples was now reduced to 534 accessions and selected as the *'FIGS Sunn pest set'*. These accessions were tested for resistance to Sunn pest at the ICARDA research station in Tal Hadya in Syria during the growth season in 2007, and the 57 most promising accessions re-tested in 2008. This pest screening experiment identified 9 accessions with good resistance to Sunn pest (El Bouhssini et al., 2009). During the last decade more than 2000 wheat accessions were screened for Sunn pest at ICARDA, under similar experimental conditions (but not using the FIGS sampling approach) and with a significantly

lower hit rate (El Bouhssini et al., 2007a *cf.* El Bouhssini et al., 2009; El Bouhssini et al., 2007b). Sunn pest is a major pest, causing severe damage in North Africa, South-East Europe, West and Central Asia.

7.1.2 Sources of wheat resistance to Russian Wheat Aphid (RWA)

The FIGS set for Russian wheat aphid (*Diuraphis noxia* Kurdjumov) was sampled in a similar manner as the Sunn pest set, and from the same set of wheat accessions from ICARDA, Vavilov Institute and AWCC (Street et al., 2008). The first sampling rule selected accessions from areas with historical known occurrence of Russian Wheat Aphid (1). The second rule excluded accessions from areas with wet climate during the growing season (2). The third rule selected for a combination of elevation and temperature (3) leaving 1125 accessions from a total of 521 distinct collecting sites chosen as the *'FIGS Russian Wheat Aphid set'*. Screening of this FIGS set identified 12 resistant accessions (El Bouhssini et al., 2010; Street et al., 2008). Russian Wheat Aphid is a global pest (except Australia) causing severe damage to wheat and barley crops (Liu et al., 2010).

7.2 Trait mining using FIGS and multivariate data analysis methods

PAPER II, III, and IV provide the results from a series of trait mining experiments using FIGS and multivariate data analysis methods. By using multivariate methods these studies provides evidence in support of the FIGS sampling approach. The following sub chapters describe the data analysis methods used and the trait mining results obtained from these studies.

7.3 Factor analysis

Many scientific experiments are conducted using automatic machines that measure a large number of properties for the samples included in a study. The resulting dataset tables are often so-called *"short and fat"*, which means that they have many more variables than they have samples (records). This often causes problems for data analysis methods that are based on the assumption of uncorrelated variables. The opportunity to make direct measurements on the biological or physical property under study is rare in such automatic registration of machine readings. Most of the properties are at best indirectly related to the property the researcher wants to explore. The researcher can choose those variables that have the closest relationship to the effect they wish to study. But when this target effect is indirectly measured it is often possible to find so-called *latent factors* as a combination of some of the recorded variables. These latent factors are thus approximations of hidden variables "inside" the dataset, and sometimes they are also good approximations of the actual effect the scientist wants to study. Principal Component Analysis (PCA) is a popular data analysis method following the principles of factor analysis. Factorial analysis provides a method to extract a more compressed representation of the variance contained in a dataset. This is mostly done in such a way that the extracted hidden variables are orthogonal to each other and thus with zero covariance. The hidden variables are called latent factors or principal components depending on the method used to extract them. In many datasets the directions in the data with the largest variation (extracted as the hidden variables) are also the directions that hold the maximum structural information. The factor analysis methods can thus be used to separate information from noise.

7.4 Data analysis methods

So-called *unsupervised* data analysis methods are used to explore the dataset to identify patterns and relationships when no class membership is assumed or no dependent (response) variables are defined. This approach can also be useful for initial explorative analysis to get to know a new dataset better. With a dependent variable(s) defined the data analysis method is a so-called a *supervised* method. When the dependent variable is on a so-called discrete measurement scale, the so-called classification or discriminant methods are usually the most appropriate data

analysis methods to use. When the dependent variable is on a continuous (or interval) scale, approaches called regression methods are usually more effective (Stevens, 1946; Velleman and Wilkinson, 1993).

7.5 Multiway data structures

A single data value can be seen as the simplest data structure. Multiple data values for one variable are called a *vector*. In multiway terminology the vector is a 1-way data structure (X_i). Multiple data values for the same objects across multiple variables are called an *array*, *matrix* or a data *table* (Figure 7.1, left). In multiway terminology this is called a 2-way data structure or a '2-way array' ($\mathbf{X}_{ij}$). The number of samples are said to define the size of first *mode* and the number of variables define the size of the second *mode*. The 3-way data structure follows from the same pattern as described from a 1-way data structure to the 2-way table. The 3-way array is constructed as repeated instances of the 2-way array for one more *mode* (Table 7.1, right). That is all the samples and all the variables of mode 2 are recorded multiple times ($\underline{\mathbf{X}}_{ijk}$). Higher-way data structures (N-way) are constructed by including multiple instances of the entire lower-level data structure. A multiway data structure (3-way and higher) is often called a *tensor*. This terminology is described in more detail by Kiers (2000). In database terminology data structures such as these N-way arrays are often called a *data cube*.

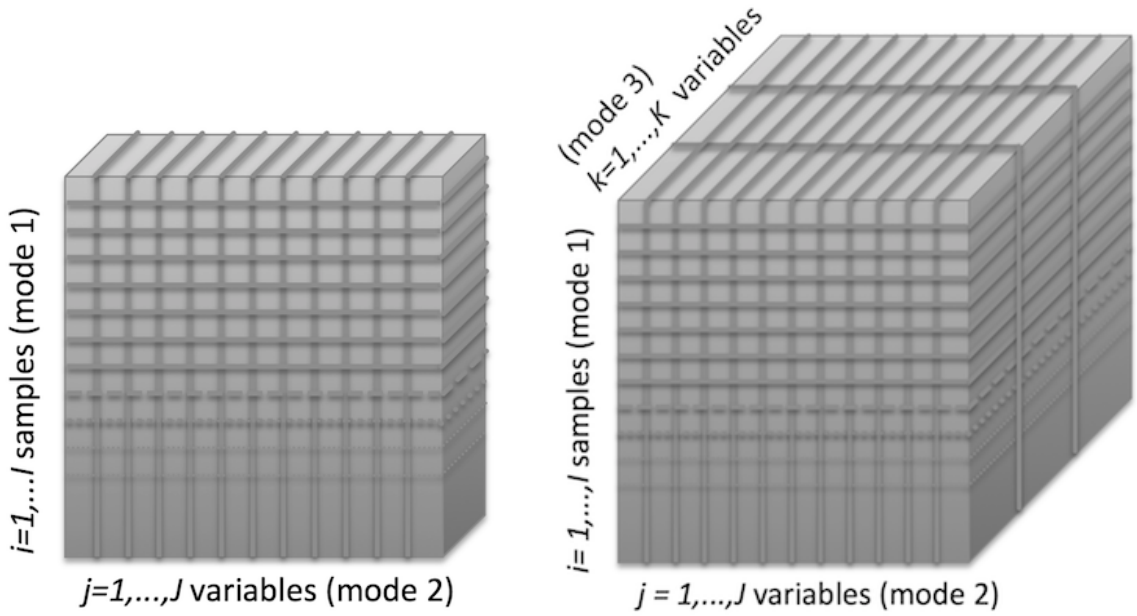


Figure 7.1: Left image: Array, matrix, 2-way ($\mathbf{X}_{ij}$). Right image: Data cube, tensor, 3-way ($\underline{\mathbf{X}}_{ijk}$)

The climate data used for the trait mining data analysis in PAPER II was organized as a 3-way array. The number of germplasm accessions was organized as the first mode. The 12 different months was organized as the second mode; and the different climate variables (with data values for each of the 12 months) were organized as the third mode. Figure 7.2 and 7.3 illustrates the transformation or so-called *'folding'* of the original source data into the 3-way data structure. First all the monthly means from the same climate variable are grouped together. Each of the climate variables is next organized "behind" each other to form a data cube. In the resulting 3-way structure each of the climate variables are called a *'slab'* or a *'slice'*. Slices can be defined horizontally, vertically and across the third mode as illustrated in the figure. The data cube can also be seen as constructed by a set of vectors, so-called *'fibers'*. Fibers can run across each of the modes and there are thus equal possible directions for *fibers* as there are modes in the data structure (Kiers, 2000).

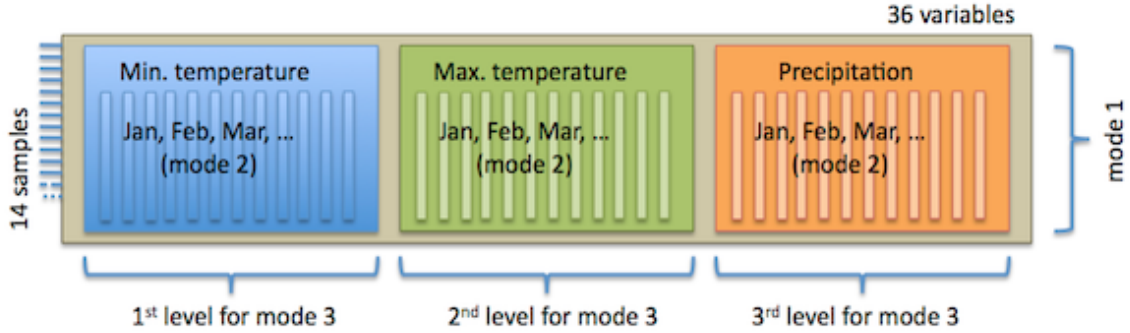


Figure 7.2: The original source climate data was organized in a data table (2-way structure).

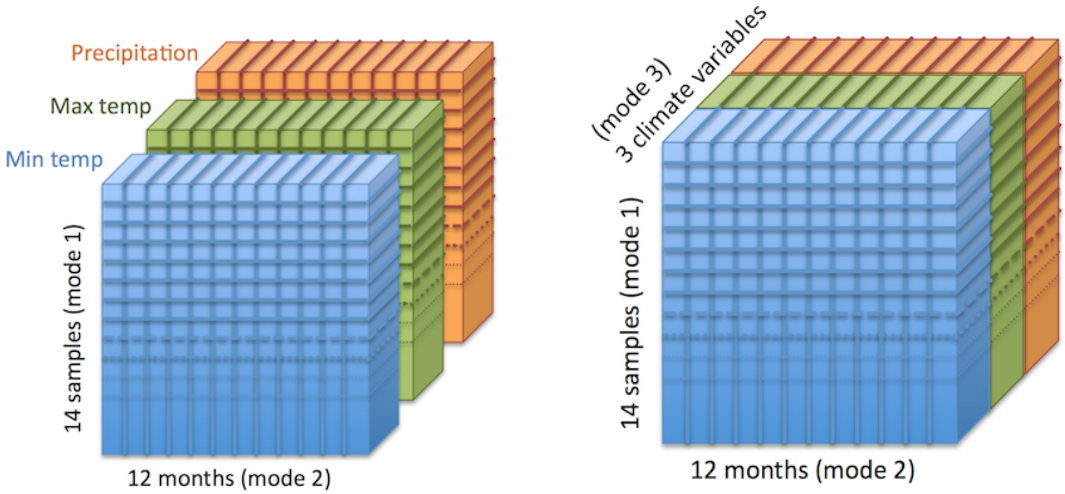


Figure 7.3: The 12 monthly means for each climate variable was organized as a 2-way *'slab'* for the 3rd mode of the 3-way data cube.

7.6 Multiway data analysis

With the dataset organized as a multiway array, systematic relationships in the data can be analyzed across each mode individually. In the 2-way array in Figure 7.2 the data analysis algorithm would analyze systematic variation for each sample across all the climate variables grouping all months together. The data analysis would not easily "see" systematic variation across the climate variables for only one month. With the 3-way structure in Figure 7.3 the "true nature" of the data is better preserved and systematic variation across individual climate variables, months and samples will be identified much more readily.

7.7 Pre-processing and cross-validation

In factorial data analysis the structural information or variance is extracted by various decompression methods. When the individual variables in a multivariate dataset have very different numerical range of data values, the variables with the highest numerical values and the largest numerical range tend to dominate the models. To give the variables a more "equal influence" on the model most datasets are treated by so-called *pre-processing* methods. Most common is the centering of data values around the mean for each variable or so-called *"mean centering"*. Trait number 3 in figure 7.4 would have influenced the model much more than other variables without mean centering. Scaling is the second most common pre-processing method, and aim to extract a more equal variance from each variable. When both centering and scaling using the inverse standard deviation as scaling factor, is applied the pre-processing method is called *"autoscaling"*. As seen in figure 7.4.b compared to figure 7.4.c the variance for trait variable 3 is reduced by the scaling method, while the variance for trait variable 1, 4 and 6 is enhanced relative to the other variables. (Source: Smilde et al., 2004:221-255).

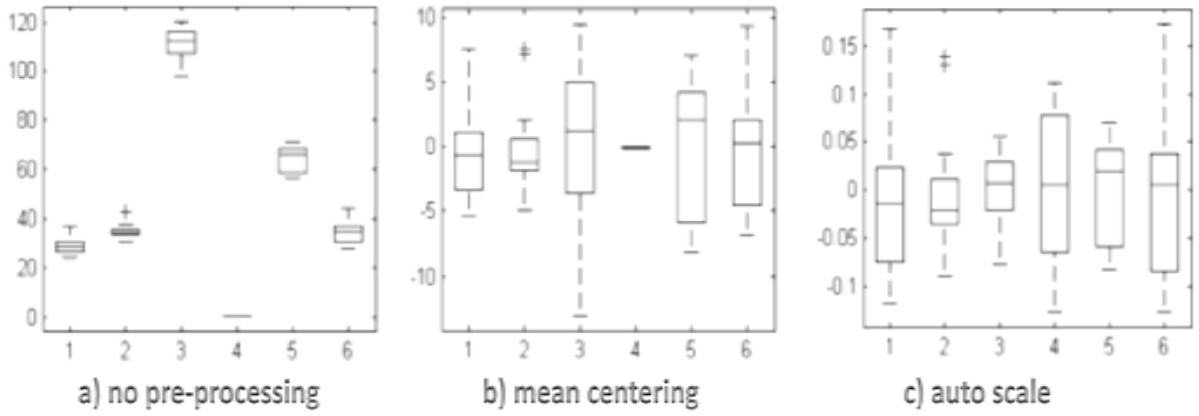


Figure 7.4: Pre-processing of the trait dataset from PAPER II, Nordic barley landraces. Traits: heading days (1), ripening days (2), plant length (3), harvest index (4), volumetric weight (5), and thousand-grain weight (6).

How well a model is fitted to the calibration dataset (training data) is not a good indicator for how suitable the model is for predicting samples not used for calibration. *"Cross-validation"* is a common approach for assessing the desired generalization of the model and to avoid specific focus on the training samples. With the *cross-validation* method some samples from the training set is kept outside during calibration process and the model is evaluated by how well these held-out samples are described by the model. Samples are held out in a systematic way so that eventually all the samples in the training set have been used both as training samples and as validation samples. A common cross-validation scheme in smaller dataset is the so-called *leave-one-out* where only one sample is held out in each calibration-validation cycle. A final set of test samples are usually held out during both the calibration and the validation phase to provide a neutral estimate of model generalization. (Source: Smilde et al., 2004:145-173).

7.8 Multiway calibration with multilinear PLS (N-PLS)

Multi-way regression using N-PLS (N-way partial least squares) has been shown to be *"superior to unfolding methods, primarily due to a stabilization of the decomposition"* (Bro, 1996:47). An example of unfolded data is illustrated by figure 7.2. The multi-way data structure often provides a more appropriate description of the dataset allowing for the analysis of relationships in the data in multiple dimensions rather than restricted to the rows and columns of 2-way data structures (arrays). N-PLS is a factorial data analysis method (7.3). The decomposition of the data to factors is made to get a more compressed description of the data that aim to isolate informative structures from noise (Varmuza and Filzmoser, 2009). Partial least squares (PLS) regression differs from principal component analysis (PCA) regression in that the decomposition of the data into factors is fitted both to the independent data (here climate data) and to the independent data (here trait data). The purpose of including the dependent data in the calibration of the model is to gain improved predictive performance. N-PLS differs from standard PLS in that the data are organized in a multi-way structure. Because the multi-way structure often provides a more appropriate description of the data, the decomposition will thus often provide a more condensed solution (Smilde et al., 2004). In other words, the number of factors included in the N-PLS model is often lower than the number of factors included in a standard PLS model. When the more condensed form is appropriate the isolation of information from noise is often more successful. The reduced number of factors also makes visual inspection of characteristic plots such as score plots, loading plots, and influence plots friendlier with fewer dimensions and thus fewer axes to examine.

Barley spikes, July 2004, by Dag Endresen

Barley (NGB19218) July 2009 by Svein Solberg

Barley, Gatersleben June 2007, by Dag Endresen

7.9 Multi-way analysis with agronomical traits for Nordic barley landraces

PAPER II presents the results from a study of *'Predictive association between trait data and ecogeographic data for Nordic barley landraces'*. This study introduces trait mining with multivariate data methods compared to the step-wise and rule-based approach that was used for previous FIGS studies. Also new in this study was the application of so-called *Multi-way* analysis methods (Smilde et al., 2004) for analysis of ecogeographic datasets. Six different morphological and agronomical trait characters measured for Nordic barley (*Hordeum vulgare* L.) landraces (Kolodinska Brantestam, 2005) were analyzed together with climate data from the online WorldClim (Hijmans et al., 2005). The climate data was organized as in figure 7.3. The trait data was here collected from one single experiment with a careful experimental design (lattice squares), replicated measurements with 2 replications, and three different trial sites and two trial years (multi-site, multi-year). The trait variables were on a continuous scale therefore the regression method was chosen. The climate variables for this study were organized as a 3-way data cube, and the N-PLS regression method (Bro, 1996) used to build a model for the correlation between the climate variables and the trait variables.

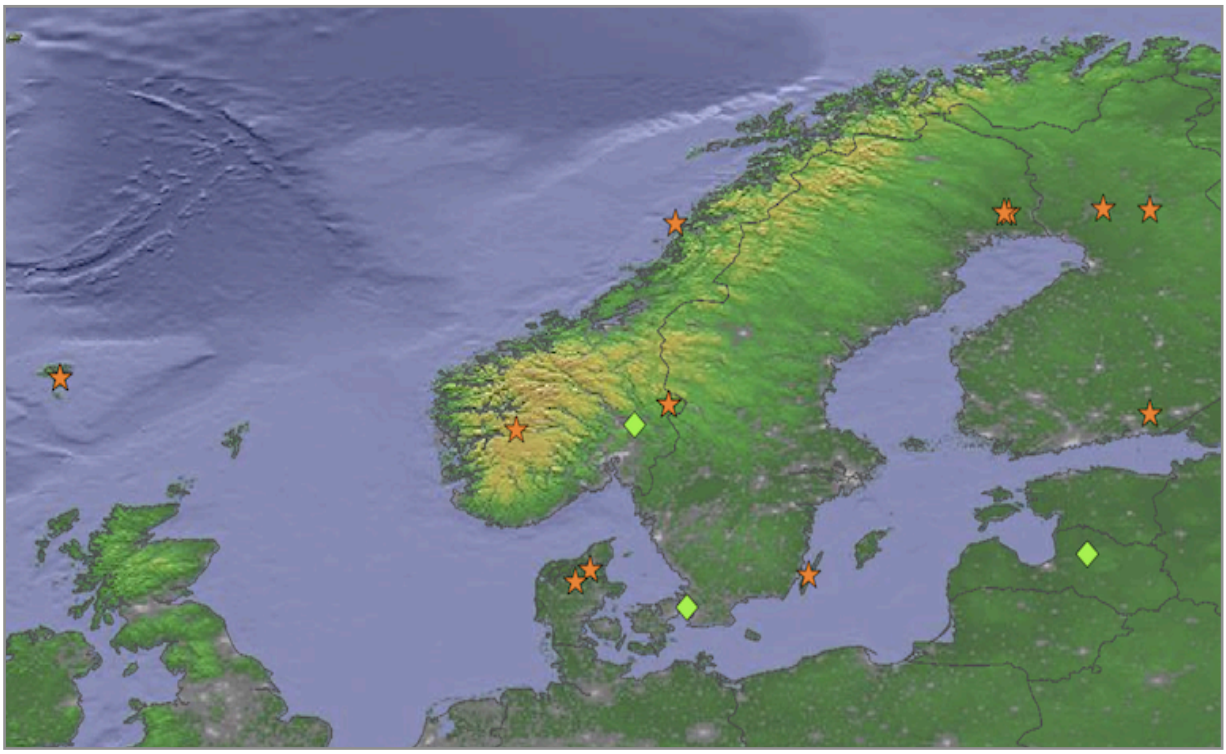


Figure 7.5: Nordic barley landraces, collecting sites are marked with orange stars, and the agricultural research stations where the field experiments were made are marked with green diamonds. Source data by Agnese Kolodinska Brantestam (2005). Map created with Quantum GIS.

The collecting site for the 14 landraces and the location of the agricultural stations where the trait observations were made (Bjørke in Norway, Landskrona in Sweden and at Priekuļi in Latvia) are displayed in figure 7.5.

A limitation for this type of studies when using Nordic germplasm is the limited number of Nordic cereal landraces that has been rescued and made available from the genebank. Crop improvement for cereal crops in the Nordic region started early and most of the landraces were replaced by high-yielding modern cultivars before the genebank conservation programs in these countries were started. In this study only 14 landrace accessions were successfully georeferenced (coordinates for the collecting site is a requirement for trait mining analysis). The computer model was calibrated with the N-PLS method and the predictive performance evaluated using cross-validation. A separate test set was not used in this experiment because the low number of samples. The multiway regression method performed notably better with regard to predictive performance than corresponding models calibrated using standard (2-way) linear regression methods such as PLS and linear regression (unpublished results). We can thus conclude that the multiway approach was successful in extracting more useful information from the dataset. One limitation encountered was the lack of effective software implementations to structure both the dependent trait variables and the independent climate variables in multiway data structures. So-called tri-PLS2 regression was explored (3-way climate data and 2-way trait data), but the predictive performance was lower than for tri-PLS1 (3-way climate data and 1-way trait data). The predictive performance could perhaps have been improved with further tuning of the model, but the visualization of the model was not very user-friendly with for example the RMSECV (root mean square error from cross validation) for each way and mode displayed together in the cross-validation plots. It is possible that the theoretical optimal data structure for the trait dataset should be a 4-way array with modes for samples (1), traits (2), trial experiment site (3) and year (4). However discouraged by the tri-PLS2 experiences such higher order models, tri-PLS4, 3-way climate data and 4-way trait data, were not explored.

Significant predictive association between the ecogeographic dataset and trait characters were identified for heading days, ripening days, plant length, harvest index, volumetric weight, but not for thousand grain weight. If there is a link between thousand-grain weight and ecogeography, these models were not able to find it. It is also possible that the thousand-grain weight is predominately selected by the farmer and not by the ecoclimate. The field trial experiment for two of the six combinations of trial location and trial year was disturbed by unusually dry May followed by unusually wet June (Priekuļi 2002) and unusually dry June (Landskrona 2003). The barley plants during these two seasons developed abnormally with respectively sprouting in the spikes or incomplete grain filling. 17 out of the remaining 20 trait datasets calibrated statistically significant trait mining models (when rejecting the Null hypothesis at the 5% probability level, p-value < 0.05). [Six model predictions were statistically significant at the 0.001 level, seven models at the 0.01 level and four models at the 0.05 level].

7.10 SIMCA (soft independent modeling of class analogies)

SIMCA (Wold, 1976; Wold and Sjostrom, 1977) is a classification method. A separate PCA model is calibrated for each class in the training dataset. The unknown samples are projected to each of these PCA models and can be classified as a member of several class models or of no class model. Figure 7.6 provides an illustration of SIMCA classification modeling.

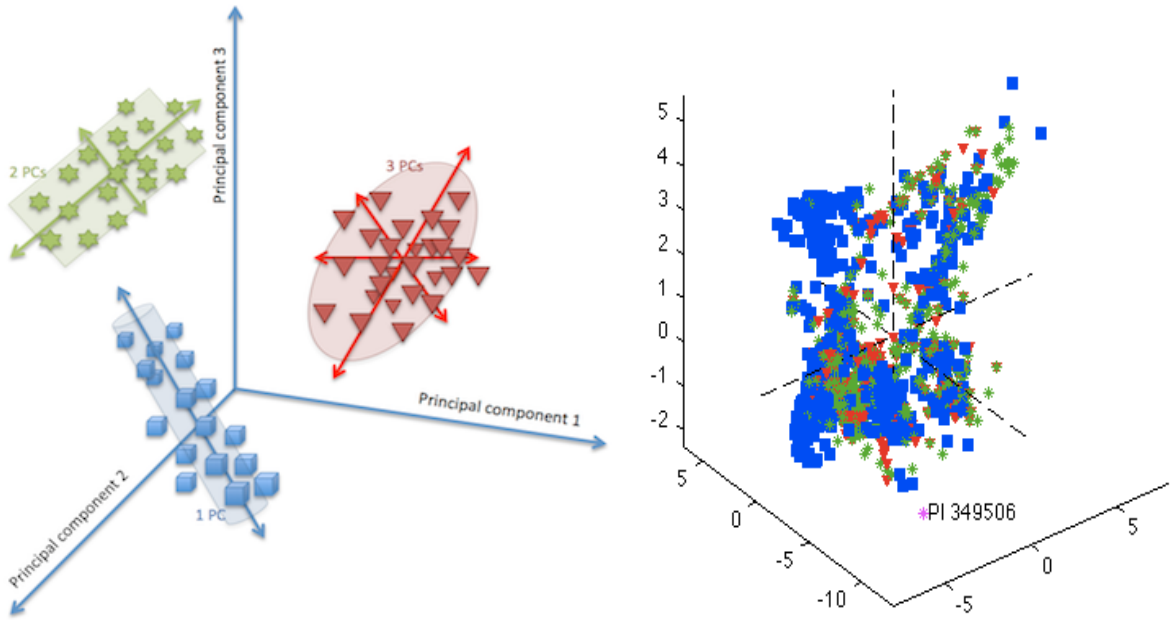


Figure 7.6: Illustrations for SIMCA classification models. **Left image**: SIMCA model. Red triangles illustrate a SIMCA PCA model with 3 principal components (PCs), green stars illustrate a PCA model with 2 PCs, and blue boxes a model with 1 PC. This figure is based on (Wise et al., 2006:201, figure 8-9). **Right image**: SIMCA model for the stem rust dataset (PAPER III). The image was created with PLS_Toolbox 5.0 (red triangles for resistant samples, green for intermediate, and blue for susceptible samples). This image is the PCA model for class 2 (intermediate samples).

7.11 Sources of resistance to fungal pathogens in wheat and barley

The second FIGS study was recently submitted for publication under the title *"Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces"* (PAPER III). The objective was to select accessions that were more likely to be resistant to a fungal pathogen. This study includes two germplasm datasets both available from the USDA NPGS GRIN database (http://www.ars-grin.gov/npgs/, verified 13 Jan 2011). Both datasets include observations for crop susceptibility to a fungal pathogen. The first dataset describes susceptibility to stem rust (*Puccinia graminis* Pers.) in bread wheat (*Triticum aestivum* L.) and durum wheat (*Triticum turgidum* L.), and the second dataset describes susceptibility to net blotch (*Pyrenophora teres* Drechs.) in barley (*Hordeum vulgare* L.). These trait observations for these fungal pathogens were recorded on an approximate interval scale as discrete numbers ranging from 0 to 9 (figure 7.7). For this type of variable scale the classification and discriminant analysis methods are the most appropriate data analysis approach to use (Stevens, 1946).

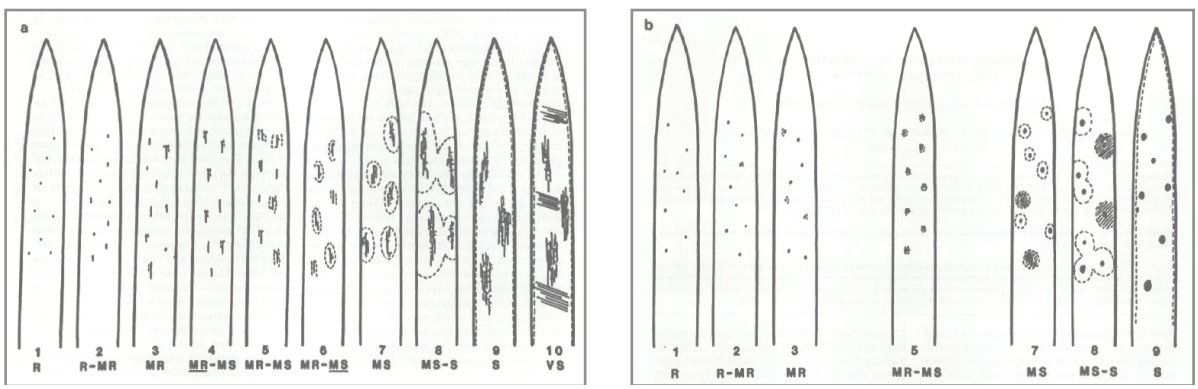


Figure 7.7: Rating of net blotch as described by Tekauz (1985:182-183). Net form left and spot form right.

7.11.1 The dataset with stem rust on wheat

The trait data with measurements of the resistance to stem rust in wheat included a total of 6889 trait scores for 4932 landraces from a total of 2013 distinct collecting sites (figure 7.7, left). The trait observations were made at two agricultural research stations located only a few km from each other in Minnesota, at Rosemount (44.72, -93.10), and at St Paul (44.99, -93.18) (figure 7.8, right).

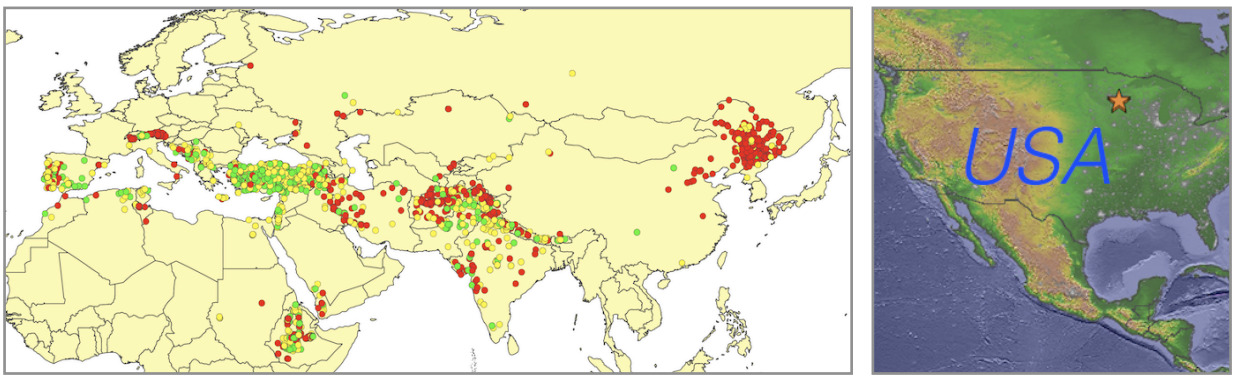


Figure 7.8: Stem rust set. **Left image**: Collecting sites for wheat landraces in the stem rust set. Latitudes span from 5.3 to 59.9; longitudes span from -9.4 to 132.8. Red dots indicate susceptible samples and green dots resistant samples. **Right image**: The stem rust screening was made at the University of Minnesota experiment stations at Rosemount and St Paul (orange stars).

7.11.2 The dataset with net blotch on barley

The trait dataset with measurements of the resistance to net blotch in barley included a total of 2786 trait scores for the same number of landraces accessions (figure 7.9, left). The trait observations were made at agricultural research stations located in Langdon in North Dakota, Stephen in Minnesota, Fargo in North Dakota, and Athens in Georgia (figure 7.9, right).

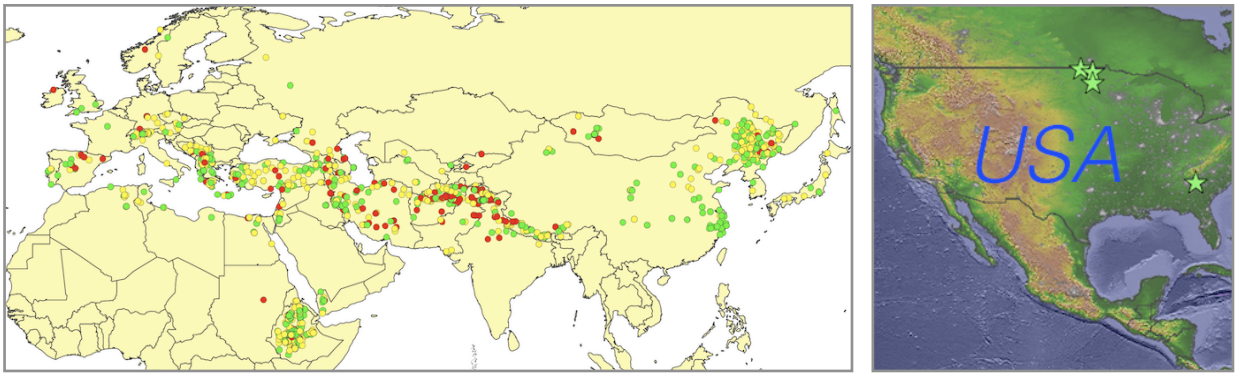


Figure 7.9: Net blotch set. **Left image**: Collecting sites for barley landraces in the net blotch set. Latitudes span from 5.3 to 66.2; longitudes span from -9.1 to 143.0. Red dots indicate susceptible samples and green dots resistant samples. **Right image**: The net blotch screening was made at the agricultural experiments stations at Langdon (ND), Stephen (MN), Fargo (ND) and Athens (GD) (green stars).

7.11.3 Data analysis results for the stem rust and net blotch set

The climate data was for this study organized as in figure 7.2. The data analysis for PAPER III was made with standard multivariate (2-way) methods including Soft independent modeling of class analogies (SIMCA; Wold and Sjostrom, 1977; Wold, 1976), Partial least squares discriminant analysis (PLS-DA; Barker and Rayens, 2003), linear discriminant analysis (LDA; Fisher, 1936), and machine learning with k-nearest neighbor (kNN; Cover and Hart, 1967). The

SIMCA method resulted in the models with the highest predictive performances. To evaluate the predictive performance the so-called positive predictive values (PPV) and the positive diagnostic likelihood ratio (LR+) were calculated (Altman and Bland, 1994). The 95% confidence intervals for each of the performance indicators were calculated. For each of the trait subsets analyzed the predictive performance of the classifiers were compared to random sampling of accessions. The upper boundary level for the 95% confidence interval of the performance indicator calculated for the random sampled set was compared to the lower boundary level for the 95% confidence interval calculated for the subset selected by each classifier. In almost all of these tests the confidence interval for the classifier was higher than and not overlapping with the confidence interval for the random sampled accessions (PAPER III). The subset in the stem rust set representing the strata for durum wheat when stratified by taxon showed marginally overlapping confidence interval with the random samples (PAPER III, Table 5). In the net blotch dataset the confidence interval for the classifier were closer to the confidence intervals for the random samples, but also here in most cases with these confidence intervals higher and not overlapping when compared to the random set. The PLS-DA classifier provided the lowest predictive performance, and generally without statistical significance as estimated by the confidence interval (PAPER III, S1).

Rust on wheat (NGB11709) by Axel Diederichsen

Winter wheat at Priekuli in Latvia, May 2004 by Bent Skovmand

Bread wheat, Lund (SE) in July 2010 by Dag Endresen

The predictive classification models were used to select landraces more likely to hold the desired trait property (resistance to the respective pathogen). All of the trait mining experiments included in this study showed a substantial improved predictive performance, both when assessed with standard classification performance indicators, and when compared to a (replicated) random sampling performed in the respective subsets. This study provides a methodology for trait mining analysis of germplasm dataset with trait data on a discrete measurement scale using the FIGS approach.

Stem rust on wheat, photo by Yue Jin, USDA Cereal Disease Lab, image K11192-1

Stem rust on wheat, photo by the Cereal Disease Lab USDA, image D939-1

USDA ARS pathologist Yue Jin evaluating stem rust on wheat, photo by USDA, image D938-1

7.12 Resistance to stem rust Ug99 in bread wheat and durum wheat

Following the positive results from the prediction of stem rust in the stem rust set from the USDA GRIN database, we set out to test this methodology for a more recent dataset with screening data on Ug99 (typified to race TTKS), which is a very virulent tribe of stem rust (*Puccinia graminis* Pers.) currently making havoc on wheat fields in East Africa and the Arabic peninsula (Singh et al., 2008). This dataset includes field experiments made in Yemen during the 2007 growth season for a total of 4563 landraces of bread wheat (*Triticum aestivum* L.) and durum wheat (*Triticum turgidum* L.). Kenneth Street from ICARDA was the project coordinator and disclosed only the Ug99 scores for a small subset of 825 landraces. The Ug99 screening scores for the remaining 3738 landraces were not disclosed to the modeler (me). This study was thus a blind test of how the method developed by the previous desktop study (PAPER III) would perform for a test more similar to a real life experiment - where the trait mining predictions will be made before the field trials. The task we assigned to the trait-mining model was to select 500 landraces more likely to be resistant to Ug99. The Ug99 dataset includes in total 10.2% resistant samples. The selected 500 accessions included 25.8% resistant samples. The subset selected by trait mining included thus 2.5 times more resistant accessions than the expected from a random sampling. This concentration of accessions for the target trait was at the same level as observed for the "desktop study" reported with PAPER III. These results are included as PAPER IV (draft manuscript).

7.13 Data analysis algorithm

The data analysis algorithm used for the experiments in PAPER III and IV is described in detail with Appendix 4. Both studies were conducted using the PLS Toolbox and MATLAB. For the model calibration and predictions the attached computer scripts were used. The SIMCA models were calibrated using a training set including 2/3 of the samples in the full trait dataset. The model complexity including the number of principal components were selected from the cross validation results using *Venetian blinds* and a split in 10 cross validation sets.

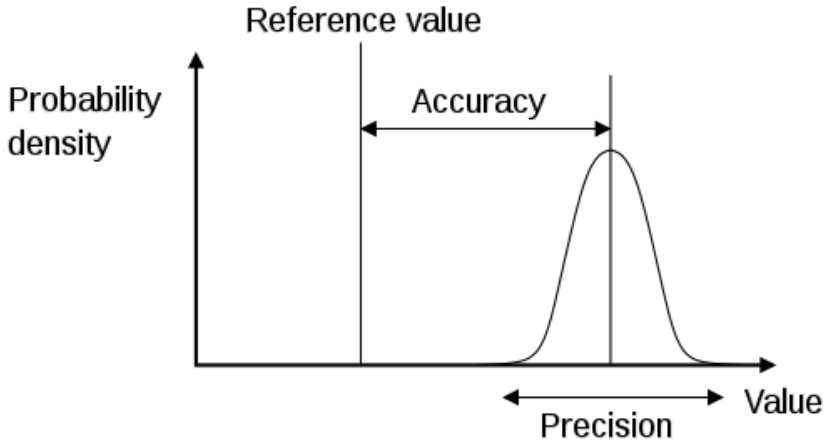


Illustration of the concepts: Accuracy and precision. By Pekaje, CC-Attrib-Alike

7.14 Representative sampling and replication in trait screening trials

The datasets analyzed with PAPER III and IV were a combination of evaluation data from several trial locations and years as were also the case for the dataset with morphological traits in Nordic barley landraces (7.9). However the stem rust and net blotch dataset did not include replications with the same accession measured at least two times. Each accession was only measured at one single trial site and trial year. Without replication it is difficult to make assessments of the accuracy (bias) and precision for the data values. Most genebank collections include a substantial number of accessions and principles of experimental design are difficult to accommodate. The reality for most genebanks is that the limitations in respect to capacity will

require a choice to be made between careful experimental designs with few accessions included, or a more rough approach with single observations and the opportunity to collect at least some evaluation data for many more accessions. The appropriate balance between a well designed replicated experiment and the desire to measure trait data for as many accessions as possible must be found. However including at least two replications provide opportunities to estimate accuracy and precision and is a highly recommended minimum principle when planning future trait screening experiments.

7.15 Conclusion

The results from PAPER II, III, and IV using the FIGS approach show that the climate layers from freely available ecogeographic databases are well suited to model and predict the reaction in the studied crops to morphological traits and to biotic stress traits. It is proposed that this approach is likely to generalize to other economically valuable crop properties. This result has the potential to improve the efficiency of field screening trials to find new sources for economically valuable crop traits.

The link between useful properties (traits) in cultivated plants including wild relatives and the ecoclimatic environment at the location of their origin was demonstrated by the agroecological studies of Vavilov (chapter 5.6; Vavilov, 1957, 1964; Dorofeev, 1992). This thesis provides experimental support for the existence of a link between the ecogeography of the collecting site and trait properties for landraces.

Trait mining with FIGS is based on the exploitation of this link for the prediction of trait properties, and the results presented by this thesis provides some of the first experimental evidence for the successful exploitation of this link to select samples with a higher concentration of a desired target trait property.

The use of trait mining in genebank collections can assist breeders to locate and utilize particular desired traits in large collections. Furthermore, trait mining with FIGS provides a new strategy to assist the breeder with more efficient identification of accessions holding target properties. The results presented with PAPER II, III and IV provide experimental support to the claim that trait mining with FIGS is an effective and efficient strategy to select samples more likely to hold a target trait property.

CHAPTER 8

8. Future work

8.1 Trait mining in wild relatives of the cultivated plants

The results presented with this thesis demonstrate that the link between ecogeographic data from the collecting site for wheat and barley landraces and trait data can be exploited with trait mining in a FIGS framework. This predictive link is assumed to be the result from adaptation in these landraces to the ecoclimatic environment where they were developed through long-term cultivation in the framework of traditional agriculture. However landraces are not only the subject of selective pressure from the environment. The farmers participate with an active role in the selection of types and properties including those more suitable for agricultural management, processing as food, or taste. In addition the environmental conditions in the farmers fields are different from those in a natural habitat with natural precipitation as the only water source (figure 1.2). These first experiments with trait mining using FIGS were conducted for landraces. It is expected that the predictive signal from the link between ecoclimate and trait properties is stronger in the wild relatives of the cultivated plants. The crop wild relatives is not the subject of selective pressures from the farmer, and is expected to express phenotypic traits more dominantly evolved in response to the natural ecoclimate. An interesting question to explore is thus if trait mining with FIGS provides a stronger predictive signal when using germplasm of crop wild relatives compared to these results using landraces.

8.2 Stratification by trait screening site and trial year

The trait experiments with data from stem rust on wheat and net blotch on barley (PAPER III) explored the effect on predictive performance from stratification of the datasets by respectively experiment location and trial year, and taxonomic levels (species and subspecies). The hypothesis was that the accessions of each stratum would be more similar in respect to the link between ecoclimate and trait property, and that the trait mining models would thus find a more accurate description of this predictive link. The results gave no clear indication of support for this hypothesis. Some of the models for subsets representing such strata calibrated models with higher predictive performance when compared to the model for the full set, but other models for strata of the same group calibrated models with lower predictive performance. It was expected to observe improved predictive performance for all (or most) models calibrated from such stratified subsets. It is possible that the lower number of samples for the subsets representing the strata compared to the higher number of samples used to calibrate the model for the full set caused a lower predictive performance in the subset models. Further studies, preferably with replicated multi-site and multi-year trait data could provide more information than the results obtained in this thesis.

8.3 Using prior trait data as the training data set

These first trait mining results indicates that the models calibrated are sensitive to the experiment conditions and that a more universal model for each target trait might be difficult to find. The study described with PAPER II revealed that the predictive performance for the same trait between trial seasons and locations variated substantially (PAPER II, Table 2). The same pattern was seen for the analysis of the stem rust dataset with PAPER III (Table 5). This indicates a limitation for using prior trait data from previous and different experimental conditions as the training set. Work has been initiated at ICARDA to guide the initial trait mining models by information provided by crop experts. If a trait evaluation experiment

continues across multiple years the trial results from the past years will provide a good training dataset that will improve the predictive performance by each subsequent trial season.

8.4 Data analysis methods

The trait mining models based on non-parametric methods provided higher predictive performances than the models based on parametric methods. Factorial methods such as SIMCA outperformed the standard linear discriminant analysis (LDA). This result suggests a non-linear link between the ecoclimatic dataset and the trait properties. Other non-linear models are thus likely to provide good predictive performances. Examples of such non-linear models include artificial neural networks (ANN; Bishop, 1996), random forest (RF; Breiman, 2001).

Some analysis methods provided very low predictive performance, but almost always at least slightly better than chance. The relatively recent data analysis approach called *'classifier ensembles'* aim to exploit the predictive performance from multiple different classifiers (Kuncheva, 2004; Rokach, 2010). The *'classifier ensemble'* is designed to extract useful predictive information from multiple, preferably fundamentally different classifiers. Such ensembles have been demonstrated to give higher combined predictive performances than even the "best" classification method included in the ensemble. The results presented with PAPER IV were combined as a simplified classifier ensemble. Each time one of the classification models (SIMCA and kNN) predicted a sample to be resistant this sample received a vote. The samples with the most votes were selected as most likely to hold this target trait property. Kuncheva (2004) and Rokach (2010) describes much more elaborate schema for the construction of classifier ensembles. This approach is expected to be effective for trait mining with FIGS, and further exploration advised. Surowiecki (2004) provides a recommended *'popular science'* presentation of related concepts.

8.4 Sampling strategies

For the experiments reported here, the trait scores were predicted by the trait mining models and ranked by the predicted score. In addition the estimated prediction probability was calculated for each of the predicted scores. This prediction probability was used as the second sorting order to arrive at a list with the desired trait properties with the highest prediction probability on the top. The top 500 samples from this list were selected as the samples most likely to hold the desired trait property. For the stem rust set the proportion resistant samples in the selected 500 samples was 2.5 times higher than the proportion in the complete set. Selecting samples directly from the top of the list ordered by predicted class and prediction probability is not the optimal sampling strategy. It is recommended to explore other sampling strategies were the subset predicted samples are sampled from the entire set, but with a strong bias towards the samples at the top of this sorted list. Such sampling strategies are known as *'probability proportional to size'* (Cochran, 1977; Wulfsohn, 2010). Probability sampling provides an assessment of the generalization of the solution obtained.

8.5 Multi-way classification methods

The multi-way regression methods provided improved predictive performance for the study on traits in Nordic barley landraces (PAPER II). The classification experiments presented with PAPER III and IV were based on standard 2-way analysis. It is likely that multi-way classification methods will give a similar improved predictive performance as it did for the regression models and a recommended topic to explore further. Parallel Factor Analysis (PARAFAC; Harshman, 1970; Carroll and Chang, 1970; Bro, 1997) can be seen as a multi-way version of PCA. PARAFAC tend to use fewer degrees of freedom and is considered a simpler, less complex model than PCA. The climate dataset can be organized as a 3-way structure as described in PAPER II. A PARAFAC model can be calibrated from the climate data and the

output (score and loading vectors) can be fitted to the trait data using regression or classification methods. The model fitted to the output from the PARAFAC model can be used to predict the trait scores. It should also be possible to make a SIMCA analysis based on PARAFAC instead of PCA to model each of the classes. I am however not aware of any software implementation that can do this.

For the Multi-way methods it would be equally interesting to try to run multilinear structures in regression against another multilinear structure! That is both trait data and climate data organized as a multi-way structure (data cube). For regression the 3-PLS3 could be suitable for a 3-way representation of the ecoclimatic data as described with figure 7.3, and a 3-way representation of the trait data with samples as the first mode, the trait variables as the second mode and the trial conditions including experiment location and trial year as the third mode. These 3-way representations of both ecoclimatic data and trait data were found to be stable in a split-half analysis on the same dataset for Nordic barley landraces as analyzed with PAPER II (unpublished results).

Literature references

Acquaah, G. (2007). Principles of plant genetics and breeding. Blackwell Publishing, Malden, USA. ISBN: 978-1-4051-3646-4.

Aldén, B. (1998). Kulturväxtlexicon. Natur och Kultur/LT, Stockholm, Sweden. 467 p. ISBN: 91-27-33907-6. [Botanical names for more than 10 000 cultivated species]

Aldén, B. and S. Ryman (2009). Våra kulturväxters namn, ursprung och användning. Forskningsrådet Formas, Formas Förlag, Stockholm, Sweden. 765 p. ISBN: 978-91-540-6026-9. [Botanical names for more than 50 000 cultivated species]

Aldén, B., S. Ryman, and M. Huldén (2005-2010). Svensk kulturväxtdatabas (SKUD), http://www.skud.se/ (offline since 1 Jan 2011, lack of funding).

Alercia, A., S. Diulgheroff, and T. Metz (2001). FAO/IPGRI Multi-crop passport descriptors, December 2001. International Plant Genetic Resources Institute (IPGRI) / Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. Available at http://apps3.fao.org/wiews/mcpd/MCPD_Dec2001_EN.pdf, verified 25 Jan 2011.

Altman, D.G. and J.M. Bland (1994). Statistical notes: Diagnostic tests 2: predictive values. BMJ 309(6947): 102.

Anderson, R.P., D. Lew, and A.T. Peterson (2003). Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecological Modeling 162: 211-232.

Anonymous (1930). Andree De Vilmorin. Journal of Heredity 21(5): 224.

Arber, W., K. Illmensee, W.J. Peacock, and P. Starlinger (eds) (1984). Genetic manipulation: Impact on man and society. Cambridge University Press, Cambridge, UK. 280 p. ISBN: 9780521264174. DOI: 10.2277/0521264170.

Arnaud, E., M. Mackay, D. Endresen, S. Dias, R. Sood, K. Viparthi, M. Skofic, A. Alercia, T. Franco, F. Atieno, X. Scheldeman, A. Shamsie, S. Louafi, and P. Cyr. (2008). A global information system for the conservation and sustainable use of plant genetic resources for food and agriculture (PGRFA). p. 78. *In*: Weitzman, A.L., and L. Belbin (eds). Proceedings of TDWG 2008. Perth, Australia. Available at http://www.tdwg.org/proceedings/article/view/396, verified 25 Jan 2011.

Bailey, L.H. (1924). Manual of cultivated plants: a flora for the identification of the most common or significant species of plants grown in the continental United States and Canada, for food, ornament, utility, and general interest, both in the open and under glass. Macmillan, New York, USA. 851 p. [Second revised edition in 1949]

Balfourier, F., G. Charmet, J.M. Prosperi, M. Goulard, and P. Monestiez (1998). Comparison of different spatial strategies for sampling a core collection of natural populations of fodder crops. Genetics Selection Evolution 30(Supplement 1): S215-S235.

Balfourier, F., V. Roussel, P. Strelchenko, F. Exbrayat-Vinson, P. Sourdille, G. Boutet, J. Koenig, C. Ravel, O. Mitrofanova, M. Beckert, and G. Charmet (2007). A worldwide bread wheat core collection arrayed in a 384-well plate. Theoretical and Applied Genetics 114: 1265-1275.

Balint-Kurti, P. and G.S. Johal (2009). Maize Disease Resistance. p. 229-250. *In*: Bennetzen, J.L. and S.C. Hake (eds). Handbook of maize: Its biology, 229. Spinger, New York, USA. ISBN: 978-0-387-79417-4. DOI: 10.1007/978-0-387-79418-1_12.

Banks, H.P. (1994). Liberty Hyde Bailey, 1858-1954, biographical memoirs. National Academy of Sciences, Washington DC, USA. Available at http://books.nap.edu/html/biomems/lbailey.pdf, verified 25 Jan 2011.

Barker, M. and W. Rayens (2003). Partial least squares for classification. Journal of Chemometrics 17(3): 166-173. DOI: 10.1002/cem.785.

Bateson, W. (1902). Mendel's Principles of Heredity, A Defense [First Edition]. Cambridge University Press, London, UK. 212 p. Available at http://www.esp.org/books/bateson/mendel/facsimile/, verified 25 Jan 2011.

Baudoin, J.P., O. Rocha, J. Degreef, A. Maquet, and L. Guarino (2004). Ecogeography, demography, diversity and conservation of *Phaseolus lunatus* L. in the Central Valley of Costa Rica. Systematic and ecogeographic studies on crop genepools 12. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. 84 p. ISBN: 92-9043-638-7.

Baur, E. (1914). *Die Bedeutung der primitiven Kulturrassen und der wilden Verwandten unserer Kulturpflanzen für die Pflanzen züchtung* [The significance of primitive cultivars and wild relatives of our crop plants for plant breeding]. Jahrbuch der Deutschen Landwirtschafts-Gesellschaft 29: 104-109.

Bellwood, P. (2005). First farmers, the origins of agricultural societies. Blackwell Publishing Ltd, Malden, MA, USA. 360 p. ISBN: 978-0-631-20566-1.

Bennett, A. (1965) Plant introduction and genetic conservation: genecological aspects of an urgent world problem. Scottish Plant Breeding Station Record, pp 17-113.

Berendsohn, W. and H. Knüpffer (2006). Draft mapping of EURISCO descriptors to ABCD 2.06 [Online]. Published online by the Berlin Botanical Garden (BGBM). Available at http://www.bgbm.org/tdwg/codata/schema/Mappings/EURISCO-2-ABCD.pdf, verified 25 Jan 2011.

Berger, D.K. (2004). Gene-mining the *Arabidopsis thaliana* genome: applications for biotechnology in Africa. South African Journal of Botany 70(1): 173-180.

Berger, J., J.A. Palta, C. Ludwig, D. Shrestha, M.C. Mackay, K.A. Street, J. Konopka, S. Jenkins, K.N. Adhikari, H.C. Clarke, J.S. Sandhu, and H. Nayyar (2008). Emerging opportunities for agriculture: investigating plant adaptation by characterizing germplasm collection habitats. *In*: Unkovich, M.J. (ed). Global Issues, Paddock Action, Proceedings of the 14th Agronomy Conference 2008, 21-25 September 2008, Adelaide, South Australia. Available at http://www.regional.org.au/au/asa/2008/concurrent/biotechnology/5781_bergerj.htm, verified 6 Jan 2011.

Berne Convention (1886). Berne Convention for the protection of literary and artistic works. Convention in Berne, Switzerland in 1886. [Latest amended version from 1979]. Available at http://www.wipo.int/treaties/en/ip/berne/index.html, verified 22 Jan 2011.

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: 9519-9524. DOI: 10.1073/pnas.0904152106.

Bhullar, N.K., M. Mackay, and B. Keller (2010a). Genetic diversity of the *Pm3* powdery mildew resistance alleles in wheat gene bank accessions as assessed by molecular markers. Diversity 2: 768-786. DOI: 10.3390/d2050768.

Bhullar, N.K., Z. Zhang, T. Wicker, and B. Keller (2010b). Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene *Pm3*: a large scale allele mining project. BMC Plant Biology 10: 88. DOI: 10.1186/1471-2229-10-88.

Biodiversity Information Standards (TDWG) [Online]. Available at http://www.tdwg.org, verified 25 Jan 2011.

BioGeomancer Working Group (2005-2007). BioGeomancer (BG) [Online]. Available at http://www.biogeomancer.org/, verified 25 Jan 2011.

Bishop, C. (1996). Neural networks for pattern recognition. Oxford University Press, UK. ISBN: 978-0198538646.

Bjarnason, S. and M. Niklasson (compilers) (1989). The Nordic barley catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-00-5. ISSN: 1100-3456. NGB publications no. 1. Available at http://www.ngb.se/Databases/Download/, verified 25 Jan 2011.

Bjørnstad, Å. (2005). Economic value of plant breeding - the missing estimate? Seminar on Nordic prebreeding programmes, Nordic Gene Bank 16th Oct 2003. Journal of the Swedish Seed Association 115(1-2): 72-77.

Blixt, S. and J.T. Williams (1982). Documentation of genetic resources: A model. Nordic Gene Bank (NGB), Alnarp, Sweden, and International Board for Plant Genetic Resources (IBPGR), Rome, Italy. 84 p. APPG:IBPGR/83/21.

Bobrov, E.G. and N.N. Tzvelev (vol eds) (2004). Flora of the USSR, alphabetical indexes to volumes 1-30. Smithsonian Institute, Washington, DC, USA, and Baba Barkhanath Printers, New Delhi, India. [Translated from Russian to English by Dr. V.S. Kothekar; scientific editor was S.W. Shetler]. Available at http://www.archive.org/details/floraofussr130bota, verified 25 Jan 2011.

Boller, B., E. Willner, L. Maggioni, and E. Lipman (2005). Report of a working group on forages. Eighth meeting, 10-12 April 2003, Linz, Austria. European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR), International Plant Genetic Resources Institute (IPGRI), Rome, Italy. ISBN: 92-9043-672-7.

Borlaug, N.E. (1954). Mexican wheat production and its role in the epidemiology of Stem rust in North America. Phytopathology 44: 398-404.

Borlaug, N.E. (2007). Sixty-two years of fighting hunger: Personal recollections. Euphytica 157: 287-297. DOI: 10.1007/s10681-007-9480-9.

Bothmer, R. von and I. Linde-Laursen (1989). Backcross to cultivated barley (*Hordeum vulgare* L.) and partial elimination of alien chromosomes. Hereditas 111(2): 145-147. DOI: 10.1111/j.1601-5223.1989.tb00388.x

Bothmer R. von, N. Jacobsen, C. Baden, R.B. Jørgensen, and I. Linde-Laursen (1995). An ecogeographic study of the genus *Hordeum*, 2nd edition. Systematic and ecogeographical studies on crop genepools 7. International Plant Genetic Resources Institute, Rome, Italy. 129 p. ISBN: 92-9043-229-2.

Brahmi, P., S. Saxena, and B.S. Dhillon (2004). The Protection of plant varieties and farmers' rights act of India. Current Science 86(3): 392-398.

Breiman, L. (2001). Random forests. Machine Learning 45(1): 5–32. DOI: 10.1023/A:1010933404324.

Bro, R. (1996). Multi-way calibration. Multi-linear PLS. Journal of Chemometrics 10(1): 47–61. DOI: 10.1002/(SICI)1099-128X(199601)10:1<47::AID-CEM400>3.0.CO;2-C.

Bro, R. (1997). PARAFAC. Tutorial and applications. Chemometrics and Intelligent Laboratory Systems 38: 149-171.

Brooks, W.K. (1883). The law of heredity. A Study of the cause of variation, and the origin of living organisms. John Murphy & Co, Baltimore, USA. Available at http://www.archive.org/details/lawofhereditystu1883broo, verified 25 Jan 2011. [Second revised edition, the first edition was published in 1876]

Brown, A.D.H. (1989a). The case for core collections. p. 136-156. *In*: Brown, A.D.H., O.H. Frankel, D.R. Marshall, and J.T. Williams (eds). The use of plant genetic resources. Cambridge University Press, Cambridge, UK. ISBN: 978-0521-368865. DOI: 10.2277/0521368863.

Brown, A.D.H. (1989b). Core collections: a practical approach to genetic resources management. Genome 31: 818-824.

Brown, A.H.D. (1995). The core collection at the crossroads. p. 3-19. *In*: Hodgkin, T., A.H.D. Brown, T.J.L. van Hintum, E.A.V. Morales (eds). Core collections of plant genetic resources. John Wiley & Sons, UK. 269 p. ISBN: 978-0-471-95545-0.

Brown, A.H.D., O.H. Frankel, D.R. Marshall, and J.T. Williams (1989). The use of plant genetic resources. Cambridge University Press, Cambridge, UK. 396 p. ISBN: 978-0521-368865. DOI: 10.2277/0521368863.

Brush, S.B. (ed) (2000). Genes in the Field: On-Farm conservation of crop diversity. International Development Research Centre (IDRC), Ottawa, Canada; International Plant Genetic Resources Institute (IPGRI), Rome, Italy; and Lewis Publishers, CRC Press, Boca Raton, FL, USA. 288 p. ISBN: 0-88936-884-8.

Busby, J.R. (1991). BIOCLIM - A bioclimatic analysis and prediction system. p. 64-68. *In* Margules, C.R., and M.P. Austin (eds). Nature Conservation: Cost effective biological surveys and data analysis. CSIRO, Canberra, Australia.

Candolle, A.P. de (1883). Origine des plantes cultivées [The origin of cultivated plants]. Librarie Germer Baillière Et Compagnie, Paris, France. 377 p. [1st edition, in French] Available at http://www.archive.org/details/originedesplant00candgoog, verified 25 Jan 2011.

Candolle, A.P. de (1884a). The origin of cultivated plants. Popular Science Monthly 25: 785-788. Available at http://www.archive.org/stream/popularsciencemo25newyuoft#page/784/mode/2up, verified 25 Jan 2011.

Candolle, A.P. de (1884b). The origin of cultivated plants. Kegan Paul, Trench, London, UK. 468 p. [1st English edition, London, 1884, http://www.archive.org/details/originofcultivat1884cand] [USA edition, 1885, http://www.archive.org/details/originofcultivat00cand, verified 25 Jan 2011].

Carroll, J.D. and J. Chang (1970). Analysis of individual differences in multidimensional scaling via N-way generalization of 'Eckart-Young' decomposition. Psychometrika 35: 283-319. DOI: 10.1007/BF02310791.

Chakravarthy, A.K. (2004). Status and spread of sugarcane woolly aphid, *Ceratovacuna lanigera* Zehntner (Hemiptera: Aphididae) in South Karnataka. Insect Environment 10: 88-90.

Chapman, A.D. and J. Wieczorek (eds) (2006). Guide to best practices for georeferencing. Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark. ISBN: 87-92020-00-3. Available from http://www2.gbif.org/BioGeomancerGuide.pdf, verified 25 Jan 2011.

Charmet, G. and F. Balfourier (1995). The use of geostatistics for sampling a core collection of perennial ryegrass populations. Genetic Resources and Crop Evolution 42: 303-309.

Chen, J. (2006). The parable of the seeds: Interpreting the plant variety protection act in furtherance of innovation policy. Notre Dame Law Review 81(4): 1-51. Minnesota Legal Studies Research Paper. Available at http://ssrn.com/abstract=784189, verified 30 Dec 2010.

Childe, V.G. (1952). New light on the most ancient east. Routledge & Paul, London, UK. 255 p. ISBN: 978-0-393-00469-4. [4th Edition; First published in 1928 under the title *'The most ancient east'*]

CHM germplasm portal (2006). Germplasm clearing house mechanism, a global portal to information and data on genetic resources [Online]. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. Available at http://chm.grinfo.net, verified 25 Jan 2011.

Christie, W.H. (1914). *Undersøkelser over norsk graaert samt nogen krydsninger mellem former av den og Pisum sativum* [Experiments with hybridization of Norwegian gray peas and *Pisum sativum*]. PhD thesis. Universitetet i Kristiania, Norway. [In Norwegian]

Cochran, W.G. (1977). Sampling Techniques, 3rd edition. Wiley, New York, USA. [*cf.* Wulfsohn, 2010]

Cohen, B.M. (1982). Recent Soviet publications. Journal of Heredity 73: 318.

Cohen, M.N. (1977). The food crisis in prehistory: Overpopulation and the origins of agriculture. Yale University Press, New Haven, CT, USA.

Correns, C. (1900). *G. Mendels Regel über das Verhalten der Nachkommenscaft der Rassenbastarde*. Berichte der Deutschen botanischen Gesellschaft 18: 158-168. [Received for publication 24 April 1900; *cf.* Roberts, 1929]

Corinna Cortes and Vladimir Vapnik (1995). Support-Vector Networks. Machine Learning 20(3): 273-297, DOI: 10.1007/BF00994018.

Cover, T.M. and P.E. Hart (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1): 21-27. DOI: 10.1109/TIT.1967.1053964.

Cribb, J. (2010). The coming famine, the global food crisis and what we can do to avoid it. University of California Press, Berkeley, USA. ISBN: 978-0-520-26071-9.

Cromarty, A.S., R.H. Ellis, and E.H. Roberts (1982). Handbooks for genebanks No 1: The Design of seed storage facilities for genetic conservation. IBPGR, Rome, Italy. 96 p. AGPE:IBPGR/82/23. Available at http://cropgenebank.sgrp.cgiar.org/images/file/learning_space/genebankmanual1.pdf, verified 17 Jan 2011.

Crow, J.F. (2001). Plant breeding giants: Burbank, the artist; Vavilov, the scientist. Genetics 158: 1391-1395.

CWR Global Portal (2007-2011). Crop wild relatives global portal [Online]. Available at http://www.cropwildrelatives.org, verified 17 Jan 2011.

Damania, A.B. (2008). History, achievements, and current status of genetic resources conservation. Agronomy Journal 100(1): 9-21. DOI: 10.2134/agrojnl2005.0239c. [Not cited in the text]

Darwin Core (DwC) [Online]. Available at http://rs.tdwg.org/dwc/, verified 25 Jan 2011.

Darwin Core extension for germplasm (DwC-G) [Online]. Available at http://rs.nordgen.org/dwc/, verified 25 Jan 2011.

Darwin, C. (1859). On the origin of species by means of natural selection, 1st edition. John Murray London, UK. Available at http://darwin-online.org.uk/contents.html#origin, verified 25 Jan 2011.

Darwin, C. (1868). The variation of animals and plants under domestication, 1st edition. John Murray, London, UK. 2 volumes. Available at http://darwin-online.org.uk/contents.html#variation, verified 25 Jan 2011.

Davies, P. (2003). An historical perspective from the Green Revolution to the Gene Revolution. Nutrition Reviews 61(6): S124-S134.DOI: 10.131/nr.2003.jun.S124-S134.

De Pauw, E. (2008). Climatic and Soil Datasets for the ICARDA Wheat Genetic Resource Collections of the Eurasia Region. Explanatory Notes. ICARDA GIS Unit, Aleppo, Syria. 68 p. Available at http://geonet.icarda.cgiar.org/geonetwork/data/regional/GRU_NetBlotch/Doc/Report_NetBlotch.pdf (6.6 MB, verified 21 Dec 2010).

Diamond, J. (1997). Guns, germs and steel. W.W. Norton, New York, USA. 480 p. ISBN: 9780099302780.

Diamond, J. (2005). Collapse. How societies choose to fail or survive. Penguin Books, London, UK. ISBN: 0-713-99286-7.

DIVA-GIS [Software]. DIVA-GIS was developed by R. Hijmans et al. Available for download from http://www.diva-gis.org, verified 25 Jan 2011

Doebley, J.F., B.S. Gaut, and B.D. Smith (2006). The molecular genetics of crop domestication. Cell 127: 1309-1321. DOI: 10.1016.j.cell.2006.12.006.

Dorofeev, V.F. (ed) (1992). N.I. Vavilov, origin and geography of cultivated plants. Cambridge University Press, Cambridge, UK. ISBN: 978-0-521-11159-1. [Translated from Russian by Doris Löve, first published in Russian in 1987]

Dorofeev, V.F., A.A. Filatenko, E.F. Migushova, R.A. Udaczin, and M.M. Jakubziner (1979). Wheat. *In*: Dorofeev, V.F. and O.N. Korovina (eds). Flora of cultivated plants, volume 1. Leningrad, Russia, Kolos. 346 p. [In Russian].

Dragavtsev, V., L. Gorbatenko, L. Bagmet, and V. Funtova (compilers) (1999). Delectus Seminum, 1999-2004. The N.I. Vavilov All-Russian Scientific Research Institute of Plant Industry (VIR), St Petersburg, Russia.

Duvick, D.N. (1984). Genetic diversity in major farm crops on the farm and in reserve, Economic Botany 38: 161-178. DOI: 10.1007/BF02858829. [Referenced by Brown, 1995]

Dworkin, S. (2009). The Viking in the wheat field. Walker & Company. 256 p. ISBN: 9780802717405.

ECPGR (2009). A strategic framework for the implementation of a European genebank integrated system (AEGIS). A policy guide. European Cooperative Programme for Plant Genetic Resources (ECPGR). Bioversity International, Rome, Italy. Available at http://aegis.cgiar.org/, http://aegis.cgiar.org/fileadmin/www.aegis.org/Documents/AEGIS_Policy_Guide.pdf, verified 25 Jan2011

ECPGR barley database [Online]. Managed by IPK Gatersleben, Germany. Available at http://barley.ipk-gatersleben.de/genres/, verified 25 Jan 2011.

ECPGR *Phleum* database [Online]. Managed by the Nordic Genetic Resources Center (NordGen). Available at http://www.nordgen.org/ecpgr/, verified 25 Jan 2011.

Edmonds, J.M. (1990). Herbarium survey of African *Corchorus* L. species. Systematic and ecogeographic studies on crop genepools 4. International Jute Organization (IJO) / IBPGR, Rome, Italy. 284 p. ISBN: 92-9043-191-1.

Edmonds, J.M. (1991). The distribution of *Hibiscus* L. section *Furcaria* in tropical East Africa. Systematic and Ecogeographic Studies on Crop Genepools 6. IBPGR, Rome, Italy. 60 p. ISBN: 92-9043-206-3.

Ehrlich, P.R. (1971). The population bomb. Ballatine Books, Cutchogue, NY, USA. 201 p.

El Bouhssini M., M. Nachit, J. Valkoun, M. Moussa, H. Ketata, O. Abdallah, M. Abdulhai, B.L. Parker, F. Rihawi, A. Joubi, and F.J. El-Haramein (2007b). Evaluation of wheat and its wild relatives for resistance to Sunn pest under artificial infestation. p. 363-368. *In*: Parker B.L., M. Skinner, M. El Bouhssini, and S.G. Kumari (eds). Sunn pest management: A decade of progress 1994-2004. Arab Society for Plant Protection, Beirut, Lebanon. 432 p. ISBN 978-9953-0-1063-2. (Information available at http://www.asplantprotection.org/SunnPest.htm, verified 11 Jan 2011).

El Bouhssini, M., B.L. Parker, M. Skinner, W. Reid, M. Nachit, J. Valkoun, M. Mosaad, O. Abdallah, A. Aw-Hassan, A. Mazid, D. Moore, S. Edgington, D. Hall, M. Maafi, R. Canhilal, M. Abdel Hay, J. El-Haramein, G. Zharmukhamedova, Z. Pulatov, and M. Dzhunusova (2007a). Integrated pest management of Sunn pest in West and Central Asia: Status and future plans. p. 61-66. *In*: Maredia, K.M., and D.N. Baributsa (eds). Integrated Pest Management in Central Asia, Proceedings of the Central Asia Region Integrated Pest Management Stakeholders Forum, Dushanbe, Tajikistan. May 27 - 29, 2007. Available at http://www.ipm.msu.edu/asia/pdf/Proceedings2007-Tajikistan.pdf, verified 11 Jan 2011.

El 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. [Published 28 April 2009]

El Bouhssini, M., K. Street, A. Amri, M. Mackay, F.C. Ogbonnaya, A. Omran, O. Abdalla, M. Baum, A. Dabbous, and F. Rihawi (2010). Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). Plant Breeding [online first]. DOI: 10.1111/j.1439-0523.2010.01814.x. [Published 6 Oct 2010]

Ellerström, S. (1982). Some notes on the origin of the Nordic Gene Bank. Journal of the Swedish Seed Association 92: 87-91.

Endersby, J. (2007). A Guinea Pig's history of biology: The plants and animals who taught us the facts of life. William Heineman, London, UK. ISBN: 9780099471240.

Endresen, D.T. (2003). SESTO user manual. Nordic Gene Bank, Alnarp, Sweden. (PHP, PostgreSQL version)

Endresen, D.T.F. (2009a). A Lifeboat to the Gene Pool, Utilization of Genetic Resources, Trait Mining with climate data. Seminar at NordGen, Alnarp, Sweden (28 April 2009, 69 slides). *Not referenced in the text.*

Endresen, D.T.F. (2009b). A Lifeboat to the gene pool, utilization of genetic resources, trait mining with climate data. LIFE KU Taastrup, Copenhagen, Denmark (Lecture for Master students, 27 May 2009, 75 slides). Available at http://www.slideshare.net/DagEndresen/, verified 25 Jan 2011. *Not referenced in the text.*

Endresen, D.T.F. (2009c). A lifeboat to the gene pool, trait mining with climate data for identification of trait properties useful for improvement of food crops. Carlsberg Research Center (CRC), Copenhagen (Seminar, 4 Nov 2009, 52 slides). Available at http://www.slideshare.net/DagEndresen/, verified 25 Jan 2011. *Not referenced in the text.*

Endresen, D.T.F. (2009d). A lifeboat to the gene pool, trait mining with eco-geographic data for identification of trait properties useful for improvement of food crops. ECPGR Cereal Pre-breeding Workshop, NordGen and SLU Alnarp, Alnarp, Sweden (25 Nov 2009, 30 slides). Available at http://www.slideshare.net/DagEndresen/, verified 25 Jan 2011. *Not referenced in the text.*

Endresen, D.T.F., B. Skovmand, and J. Bäckman (2005a). Integrated generic regional genetic resources information system. *In*: ASA-CSSA-SSSA International Annual Meeting, Salt Lake City, 5-11

November 2005. Available at http://crops.confex.com/crops/2005am/techprogram/P6077.HTM, verified 24 Jan 2011.

Endresen, D.T.F., V. Kukk, and R. Baltrenas (2005b). Regional Nordic-Baltic database cooperation. Nordic Gene Resources, livestock, crops, forest trees 4: 15.

Endresen, D.T.F., M. Mackay, and K. Street (2007). Utilization of Genetic Resources, Prediction of agricultural traits in plant genetic resources with ecological parameters. Vavilov seminar, 13 June 2007 at IPK Gatersleben, Gatersleben, Germany (43 slides). Available at http://www.slideshare.net/DagEndresen/, verified 25 Jan 2011. *Not referenced in the text.*

Engels, J.M.M., L. Maggioni, N. Maxted, and M.E. Dulloo (2008). Complementing *in situ* conservation with *ex situ* measures. p. 169-181. *In*: Iriondo, J.M., N. Maxted, and M.E. Dullo. Conserving plant genetic diversity in protected areas. CABI, Wallingford, Oxfordshire, UK. ISBN: 978-1-84593-282-4.

Enneking, D., E. Schliephake, and H. Knüpffer (2003). Enhancing the practical value of barley genetic resources in Europe through evaluation and documentation. From biodiversity to genomics: Breeding strategies for small grain cereals in the third millennium. p. 15-17. *In*: Proceedings Eucarpia Cereal Section, Salsomaggiore, Italy. Fiorenzuola d'Arda: Experimental Institute for Cereal Research.

Esakov, V.D. (compiler) (1980). *Nauchnoe Nasledstvo: Tom Piatyi. Nikolai Ivanovich, iz epistoliarnovo nasledиia, 1911-1928* [The Scientific inheritance: Volume 5. Nikolai Ivanovich, from the epistolary legacy, Vavilov's letters 1911-1928]. Nauka Press [Science Press], Moscow, USSR. 428 p. [In Russian] [See review by Cohen (1982) in Heredity]

Eshbaugh, W.H. (1993). Peppers: history and exploitation of a serendipitous new crop discovery. p. 132–139. *In*: Janick, J. and J.E. Simon (eds). New crops. Wiley, New York. Available at http://www.hort.purdue.edu/newcrop/proceedings1993/v2-132.html, verified 25 Jan 2011.

Estabrook, G.F. and R. Brill (1969). Theory of the TAXIR accessioner. Mathematical Biosciences 5: 327-340. [*cf*. Hersh and Rogers, 1975]

EURISCO (2003). EPGRIS final meeting. 11-13 September 2003, Prague, Czech Republic. PGR documentation and information in Europe. Towards a sustainable and user-oriented information infrastructure. Conference report available online at http://www.ecpgr.cgiar.org/Networks/Info_doc/FinalMeetingReports.htm, verified 25 Jan 2011.

EURISCO (2003-2011). European Search Catalogue for Plant Genetic Resources [Online database]. ECPGR, Rome, Italy. Available from http://eurisco.ecpgr.org, verified 25 Jan 2011.

Evans, L.T. (1999). Sir Otto Herzberg Frankel. 4 November 1900 - 21 November 1998: Elected F.R.S. 1953. Biographical Memoirs of Fellows of the Royal Society 45: 165-181. DOI: 10.1098/rsbm.1999.0012.

Ewens, W.J. (1972). The sampling theory of selectively neutral alleles. Theoretical Population Biology, 3(1): 87-112. DOI: 10.1016/0040-5809(72)90035-4.

FAO (1983). The international undertaking on plant genetic resources for food and agriculture. FAO, Rome, Italy. Available at http://www.fao.org/ag//CGRFA/iu.htm, verified 25 Jan 2011.

FAO (1996a). State of the world's plant genetic resources for food and agriculture, Rome, Italy. [Draft version presented at the Leipzig Conference]

FAO (1996b). Global plan of action for the conservation and sustainable utilization of plant genetic resources for food and agriculture. Rome, Italy. Available at http://www.globalplanofaction.org/, verified 29 Dec 2010.

FAO (1997). State of the world's plant genetic resources for food and agriculture, Rome, Italy. Available at ftp://ftp.fao.org/docrep/fao/meeting/015/w7324e.pdf, verified 25 Jan 2011.

FAO (2001). Press release 01/81 C5. International treaty on plant genetic resources for food and agriculture approved by FAO conference [Online]. FAO, Rome, Italy. Available at http://www.fao.org/WAICENT/OIS/PRESS_NE/PRESSENG/2001/pren0181.htm, verified 29 Dec 2010.

FAO (2002). International treaty on plant genetic resources for food and agriculture (ITPGRFA). FAO, Rome, Italy. 45 p. Available at http://www.planttreaty.org/, verified 25 Jan 2011.

FAO (2009). The international treaty on genetic resources for food and agriculture (ITPGRFA) [edition 2009]. FAO, Rome, Italy. Available at http://www.fao.org/docrep/011/i0510e/i0510e00.HTM, verified 29 Dec 2010.

FAO (2010). The second report on the state of the world's plant genetic resources for food and agriculture. Commission on Genetic Resources for Food and Agriculture (CGRFA), Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. ISBN: 978-92-5-106534-1. Available at http://www.fao.org/docrep/013/i1500e/i1500e00.htm, verified 25 Jan 2011.

FAO and ITPGRFA (2008). Information management in support of the global system for the conservation and sustainable use of plant genetic resources for food and agriculture - Global Information on Germplasm (GIG). Report from the second technical consultation on information technology support for the implementation of the Multilateral System of access and benefit sharing of the International Treaty, Rome, 2-3 December 2008. FAO, Rome, Italy. IT/GB-2/07/REPORT. (Principal investigators: Michael Mackay and Elizabeth Arnaud). Available at ftp://ftp.fao.org/ag/agp/planttreaty/gb3/tcit2/tcit2i02.pdf, verified 25 Jan 2011.

Feuillet, C., P. Langridge, and R. Waugh (2007). Cereal breeding takes a walk on the wild side. TRENDS in Genetics 24(1): 24-32. DOI: 10.1016/j.tig.2007.11.001.

Fisher, R.A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. Philosophical Transactions of the Royal Society of Edinburgh 52: 399-433. Available at http://www.library.adelaide.edu.au/digitised/fisher/9.pdf, verified 2 Jan 2011.

Flood, J. (2010). The importance of plant health to food security. Food Security 2(3): 215-231. DOI: 10.1007/s12571-010-0072-5.

Flyaksberger K.A. (1915a). *Opredelitel' pshenits* [Identification key of wheats]. Bulletin of the Bureau of Applied Botany 8(1/2): 3-190. [In Russian]

Flyaksberger K.A. (1915b). *Obzor raznovidnostey pshenitsy Sibiri* [Survey of botanical varieties of wheat in Siberia]. Bulletin of the Bureau of Applied Botany 8(8): 557-862. [In Russian]

Flyaksberger K.A. (1935). Flora of cultivated plants, Volume I, Cereals: Wheat. Kolos, Moscow-Leningrad, USSR. [In Russian]

Fowler, C. (2010). Bubbles [Online] Crop Diversity Topics, Newsletter No 24, Dec 2010. Global Crop Diversity Trust, Rome, Italy. Available at http://www.croptrust.org/documents/newsletter/newsletter_croptrust_v24.htm, verified 22 Jan 2011.

Fowler, C., and P. Mooney (1990) Shattering: Food, politics, and the loss of genetic diversity. University of Arizona Press, Tucson, USA. ISBN: 0-8165-1181-0.

Fowler, C., G. Moore, and G.C. Hawtin (2003). The international treaty on plant genetic resources for food and agriculture: A primer for the Future Harvest centres of the CGIAR. International Plant Genetic Resources Institute, Rome, Italy. Available at http://www.sgrp.cgiar.org/sites/default/files/SgrpTreaty_final.pdf, verified 25 Jan 2011.

Fowler, C., W. George, H. Shands, B. Skovmand, M. Qvenild, and G. Hawtin (2004). Study to assess the feasibility of establishing a Svalbard Arctic seed depository for the international community. Prepared for the Ministry of Foreign Affairs and the Ministry of Agriculture. Center for International Environment and Development Studies (NorAgric), Ås, Oslo and the Nordic Gene Bank (NGB), Alnarp, Sweden. Available at http://www.croptrust.org/documents/Svalbard%20feasibility%20study.pdf, and at http://www.regjeringen.no/Upload/LMD/kampanjeSvalbard/Vedlegg/Frohvelv_Study_to_assess.pdf verified 25 Jan 2011.

Franco, J., J. Crossa, S. Taba, and H. Shands (2005). A sampling strategy for conserving genetic diversity when forming core subsets. Crop Science 45: 1035-1044. DOI: 10.2135/cropsci2004.0292.

Frankel, O.H. (1950). The development and maintenance of superior genetic stocks. Heredity 4: 89-102. DOI: 10.1038/hdy.1950.6.

Frankel, O.H. (1977). Natural variation and its conservation. p. 21-24. *In:* Muhammed, A., and R.C. von Bostel (eds). Genetic diversity of plants. Plenum Press, University of Michigan, USA. ISBN: 9780306365089.

Frankel, O.H. (1984). Genetic perspectives of germplasm conservation. p. 161-170. *In*: Arber, W., K. Illmensee, W.J. Peacock, P. Starlinger (eds). Genetic manipulation: Impact on man and society. Cambridge University Press, Cambridge, UK. 280 p. ISBN: 9780521264174. DOI: 10.2277/0521264170.

Frankel, O.H. (1987). Genetic resources: the founding years. Part four: after twenty years. Diversity 11: 25-27.

Frankel, O.H. (1988). Nikolai Ivanovich Vavilov 1887-1943 a memoir. PGR Newsletter 72: 1-2.

Frankel, O.H. (1990). The future of the global genetic resources network: Activation or dissolution? Diversity 6: 59-60.

Frankel, O.H. and A.H.D. Brown (1984a). Current plant genetic resources: A critical appraisal. p. 1-11. *In:* Chopra V.L., B.C. Joshi, R.P. Sharma, and H.C. Bansal (eds). Genetics, new frontiers: Proceedings of the XV international congress on genetics. Volume 4: Applied Genetics. IBH Publishing Co, New Delhi, India. 309 p. ISBN: 0890590389.

Frankel, O.H., and A.H.D. Brown (1984b). Plant genetic resources today: a critical appraisal. p. 249-257. *In*: Holden, J.H.W., and J.T. Williams (eds). Crop genetic resources: Conservation & evaluation. George Allen & Unwin, London, UK. 296 p. ISBN: 0045810184.

Frankel O.H. and E. Bennett (eds) (1970). Genetic resources in plants - their exploration and conservation. IBP Handbook No 11. Blackwell, Oxford, UK. 554 p. ISBN: 632-05730-0.

Frankel, O.H. and J.G. Hawkes (eds) (1975). Crop genetic resources for today and tomorrow. International Biological Programme (IBP) 2. Cambridge University Press, Cambridge, UK. ISBN 0-521-20575-1.

Franklin, J. (2010). Mapping species distributions. Spatial inference and prediction. Cambridge University Press, Cambridge, UK. ISBN: 978-0-521-70002-3.

Fujisaka, S., D. Williams, and M. Halewood (eds) (2009). The impact of climate change on countries' interdependence on genetic resources for food and agriculture. Background study paper No.48. FAO Commission on Genetic Resources for Food and Agriculture. FAO, Rome. Available at ftp://ftp.fao.org/docrep/fao/meeting/017/ak532e.pdf, verified 29 Dec 2010.

Fujisaka, S., D. Williams, and M. Halewood (2010). The impact of climate change on countries' interdependence on genetic resources for food and agriculture. An executive summary. Bioversity International, Rome, Italy. 4 p. Available at http://agrobiodiversityplatform.org/climatechange/files/2010/12/11_WEB.pdf, verified 20 Jan 2011.

Galton, F. (1865). Hereditary talent and character. Macmillan's Magazine 12: 157-166, 318-327. Available at http://galton.org/essays/1860-1869/galton-1865-hereditary-talent.pdf, verified 25 Jan 2011.

Galton, F. (1871a). Experiments in pangenesis, by breeding from rabbits of a pure variety, into whose circulation blood taken from other varieties had previously been largely transfused. Proceedings of the Royal Society 19: 393-410. Available at http://galton.org/essays/1870-1879/galton-1871-roy-soc-pangenesis.pdf, verified 25 Jan 2011.

Galton, F. (1871b). Pangenesis. Nature 4: 5-6. Available at http://galton.org/essays/1870-1879/galton-1871-nature-darwin-pangenesis.pdf, verified 25 Jan 2011.

Galton, F. (1875). A theory of heredity. Contemporary Review 27: 80-95. Available at http://galton.org/essays/1870-1879/galton-1875-jaigi-theory-heredity.pdf, verified 2 Jan 2011.

Galton, F. (1876). A theory of heredity. Journal of the Anthropological Institute 5: 329–348. Available at http://galton.org/essays/1870-1879/galton-1875-cont-rev-theory-heredity.pdf, verified 2 Jan 2011.

Ganeshaiah, K.N, N. Barve, N. Nath, K. Chandrashekara, M. Swamy, and R.U. Shaanker (2003). Predicting the potential geographical distribution of the sugarcane woolly aphid using Garp and DIVA-GIS. Current Science 85(11): 1526-1528. Available at http://eprints.atree.org/36/, verified 25 Jan 2011.

GBIF Integrated Publishing Toolkit (IPT) [Online Software]. Available at http://ipt.gbif.org/ and at http://code.google.com/p/gbif-providertoolkit/, verified 25 Jan 2011.

GeoNetwork [Software] Open Source Geospatial Foundation. Available at http://geonetwork-opensource.org/, verified 25 Jan 2011.

Gepts, P. (2006) Plant genetic resources conservation and utilization: the accomplishments and future of a societal insurance policy. Crop Science 46: 2278–2292.

Geric, I., M. Zlokolica, C. Geric, and C.W. Stuber (1989). Races and populations of maize in Yugoslavia: Isozyme variation and genetic diversity. Systematic and Ecogeographic Studies on Crop Genepools 3. IBPGR, Rome, Italy. 108 p. ISBN: 92-9043-185-7.

Global Biodiversity Information Facility (GBIF) [Online]. Available at http://www.gbif.org, verified 25 Jan 2011.

Google, Inc. [Online]. Google Maps. Available at http://maps.google.com, verified 25 Jan 2011. Google Inc., TerraMetrics Inc. and Tele Atlas BV. Mountain View, CA.

Gouesnard, B., T.M. Bataillon, G. Decoux, C. Rozale, D.J. Schoen, and J.L. David (2001). MSTRAT: An algorithm for building germ plasm core collections by maximizing allelic or phenotypic richness. Journal of Heredity 92(1): 93-94. [MSTRAT is available for download from http://www.ensam.inra.fr/gap/MSTRAT/mstratno.htm, verified 25 Jan 2011]

Govindasamy, B., P.B. Duffy, and J. Coquard (2003). High-resolution simulations of global climate, part 2: effects of increased greenhouse cases. Climate Dynamics 21: 391-404. DOI: 10.1007/s00382-003-0339-z.

GRASS GIS (1982-2011). Geographic Resources Analysis Support System [Software]. Available for download from http://grass.osgeo.org/, verified 25 Jan 2011.

GRDC (2002-2009). Research, development & extension investment portfolio [Online]. Grains Research and Development Corporation (GRDC), Barton, *Australian Capital Territory (*ACT), Australia. Available at http://www.grdc.com.au/director/about/investmentportfolio, verified 25 Jan 2011.

GRDC (2004-2009). GRDC Annual Report [Online]. Grains Research and Development Corporation (GRDC), Barton, *Australian Capital Territory (*ACT), Australia. Available at http://www.grdc.com.au/director/about/corporategovernance/Annual%20Report, verified 11 Jan 2011.

Guarino, L., V. Ramanatha Rao, and V.R. Reid (eds) (1995). Collecting plant genetic diversity. Technical guidelines. CAB International, Wallingford, UK. 750 p. ISBN: 9780851989648.

Guarino, L., N. Maxted, and E.A. Chiwona (eds) (2005). A methodological model for ecogeographic surveys of crops. IPGRI Technical Bulletin No. 9. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. 55 p. ISBN: 92-9043-690-5.

Guralnick R.P., J. Wieczorek, R. Beaman, and R.J. Hijmans (2006). BioGeomancer: Automated georeferencing to map the world's biodiversity data. PLoS Biology 4(11): e381. DOI: 10.1371/journal.pbio.0040381.

Hamblin, M.T., T.J. Close, P.R. Bhat, S. Chao, J.G. Kling, K.J. Abraham, T. Blake, W.S. Brooks, B. Cooper, C.A. Griffey, P.M. Hayes, D.J. Hole, R.D. Horsley, D.E. Obert, K.P. Smith, S.E. Ullrich, G.J. Muehlbauer, and J.L. Jannink (2010). Population structure and linkage disequilibrium in U.S. barley germplasm: Implications for association mapping. Crop Science 50(2): 556-566. DOI: 10.2135/cropsci2009.04.0198.

Hammer, K., and A. Diederichsen (2009). Evolution, status and perspectives for landraces in Europe. p. 23-44. *In*: Veteläinen, M., V. Negri, and N. Maxted (eds). European landraces: on-farm conservation, management and use. Bioversity International, Rome, Italy. 344 p. ISBN: 978-92-9043-805-2.

Hanelt, P. and IPK (eds) (2001). Mansfeld's Encyclopedia of Agricultural and Horticultural Crops (Except Ornamentals). 6 volumes, 3716 p. IPK Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, and Springer, Berlin, Germany. ISBN: 3-540-41017-1. [Includes 6121 cultivated plant species]

Harlan, H.V. (1957). One Man's Life With Barley. Exposition Press, New York, USA.

Harlan, H.V. and M.L. Martini (1936). Problems and results of barley plant breeding. p. 303-346. *In*: USDA yearbook of agriculture. US Government Printing Office, Washington, DC.

Harlan, J.R. (1972). Genetic resources in *Sorghum*. p. 1-13. *In*: Rao, N.G.P., and L.R. House (eds). Sorghum in the seventies. Proceedings International Symposium on Sorghum, 27-30 Oct 1971, Hyderabad, India. Oxford and IBH, New Delhi, India. 638 p.

Harlan, J.R. (1975). Crops & Man. American Society of Agronomy, Madison, WI, USA. ISBN: 089118032X.

Harlan, J.R. (1975). Our vanishing genetic resources. Science 188: 618-621. DOI: 10.1126/science.188.4188.617.

Harlan, J.R. (1992). Crops and Man. Second edition. American Society of Agronomy, Madison, WI, USA. ISBN: 0-89118-107-5. [First published in 1975]

Harlan, J.R. (1995). The living fields: Our agricultural heritage. Cambridge University Press, Cambridge, UK. ISBN: 0-521-64992-7.

Harshman, R.A. (1970). Foundations of the PARAFAC procedure: models and conditions for an 'explanatory' multi-modal factor analysis. UCLA working papers in phonetics 16: 1-84.

Hatirli, S.A., B. Ozkan, and C. Fert (2005). An econometric analysis of energy input-output in Turkish agriculture. Renewable and Sustainable Energy Reviews 9(6): 608-623. DOI: 10.1016/j.rser.2004.07.001.

Hawkes, J.G. (1988). N.I. Vavilov the man and his work. PGR Newsletter 72: 3-5.

Hawtin, G. (2004). The Svalbard International Seed Vault: Technical, administrative and policy considerations. A report to the Government of Norway. Manor Farm House, Dorset, UK.

Hazekamp, T., J. Serwiński, and A. Alercia (1997) *In*: Lipman, E., M.W.M. Jongen, Th.J.L. van Hintum, T. Grass, and L. Maggioni (eds) (1997). Central crop databases: Tools for plant genetic resources management. International Plant Genetic Resources Institute, Rome, Italy/CGN, Wageningen, Netherlands. ISBN 92-9043-320-5.

Heidari, M.D. and M. Omid (2011). Energy use patterns and econometric models of major greenhouse vegetable productions in Iran. Energy 36(1): 220-225. DOI: 10.1016/j.energy.2010.10.048.

Herborg, L.M, C.L. Jerde, D.M. Lodge, G.M. Ruiz, and H.J. Macisaac (2007). Predicting invasion risk using measures of introduction effort and environmental niche models. Ecological Applications 17(3): 663-674. DOI: 10.1890/06-0239.

Hersh, G.N. and D.J. Rogers (1975). Documentation and information requirements for genetic resources application. Pages 407-446. *In*: Frankel, O.H. and J.G. Hawkes (eds) (1975). Crop genetic resources for today and tomorrow. International Biological Programme (IBP) 2. Cambridge University Press, Cambridge, UK. ISBN 0-521-20575-1.

Hijmans, R.J., L. Guarino, M. Cruz, and E. Rojas (2001). Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. Plant Genetic Resources Newsletter 127: 15-19. [DIVA-GIS is available at http://www.diva-gis.org/, verified 25 Jan 2011]

Hijmans, R.J., D.M. Spooner, A.R. Salas, L. Guarino, and J. de la Cruz (2002). Atlas of wild potatoes. Systematic and ecogeographic studies on crop genepools No. 10. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. ISBN: 92-9043-518-6.

Hijmans, R.J., M. Jacobs, J.B. Bamberg, and D.M. Spooner (2003). Frost tolerance in wild potato species: Assessing the predictivity of taxonomic, geographic, and ecological factors. Euphytica 130: 47-59.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25(15): 1965-1978. DOI: 10.1002/joc.1276. [The WorldClim GIS layers are available from http://worldclim.org, verified 25 Jan 2011]

Hill, A.W. (1915). The history and functions of botanic gardens. Ann. Missouri Botanical Garden, 2: 185-240. Available at http://www.archive.org/details/historyfunctions00hilluoft, verified 2 Jan 2011.

Hintum T.J.L. van (1999). The general methodology for creating a core collection. p. 10-17. *In*: Johnson R.C., and T. Hodgkin (eds). Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome. ISBN: 92-9043-424-4.

Hintum, Th.J.L. van and H. Knüpffer (1995). Duplication within and between germplasm collections. 1. Identification of duplication on the basis of passport data. Genetic Resources and Crop Evolution 42:127-133.

Hintum, Th.J.L. von, R. von Bothmer, G. Fischbeck, and H. Knüpffer (1990). The establishment of the Barley Core Collection. Barley Newsletter 34: 41-42.

Hintum, T.J.L. van, L. Frese, and P.M. Perret (1991). Crop networks, searching for new concepts for collaborative genetic resources management. Papers of the EUCARPIA/IBPGR symposium held in Wageningen, The Netherlands, 3-6 December 1990. International crop network series 4. International Board for Plant Genetic Resources (IBPGR), Rome, Italy. ISBN: 92-9043-130-X.

Hintum, Th.J.L. van, R. von Bothmer, and D.L. Visser (1995). Sampling strategies for composing a core collection of cultivated barley (*Hordeum vulgare* s. lat.) collected in China. Hereditas 122: 7-17. DOI: 10.1111/j.1601-5223.1995.00007.x.

Hintum, Th.J.L. van, A.H.D. Brown, C. Spillane, and T. Hodgkin (2000). Core collections of plant genetic resources. IPGRI technical bulletin No. 3. International Plant Genetic Resources Institute, Rome, Italy. 48 p. ISBN: 978-92-9043-454-2.

Hodgkin, T. (1991). The core collection concept. p. 43-48. *In*: van Hintum, Th.J.L., L. Frese, and P.M. Perret. Crop Networks. Searching for new concepts for collaborative genetic resources management. Papers of the EUCARPIA/IBPGR symposium held in Wageningen, The Netherlands, 3-6 December 1990. International Crop Network Series No 4. International Board for Plant Genetic Resources, Rome, Italy. ISBN: 92-9043-130-X.

Hodgkin, T., A.H.D. Brown, T.J.L. van Hintum, and E.A.V. Morales (eds) (1995). Core collections of plant genetic resources. John Wiley & Sons, UK. 269 p. ISBN: 978-0-471-95545-0.

Hoisington, D., M. Khairallah, T. Reeves, J.M. Ribaut, B. Skovmand, S. Taba, and M. Warburton. (1999). Plant genetic resources: What can they contribute toward increased crop productivity? PNAS 96: 5937-5943.

Holbrook, C.C. and W. Dong (2005). Development and evaluation of a Mini Core collection for the U.S. Peanut Germplasm Collection. Crop Science 45: 1540-1544. DOI: 10.2135/cropsci2004.0368.

Holden, J.H.W. (1984). The second 10 years. p. 277-285. *In*: Holden J.H.W., and J.T. Williams (eds). Crop genetic resources: Conservation and evaluation. George Allen and Unwin, London, UK. ISBN: 0-04-581018-4.

Holden, J.H.W. and J.T. Williams (eds) (1984). Crop genetic resources: Conservation & evaluation. George Allen & Unwin, London, UK. 296 p. ISBN: 0045810184.

Hort, A.W. (1916). Theophrastus Enquiry into plants and minor works on odours and weather sign. In two volumes. William Heinemann, London, UK. Available at http://www.archive.org/details/enquiryintoplant01theouoft [volume 1] and http://www.archive.org/details/enquiryintoplant02theouoft [volume 2] verified 2 Jan 2011.

Huldén, M. (1997). European *Phleum* database. *In*: IPGRI Newsletter for Europe 10: 9.

Huldén, M. (1999). Development of the information system at the NGB 1989-1999. p. 26-28 *In*: Nordic Gene Bank 1979 - 1999. Nordic Gene Bank, Alnarp, Sweden. ISRN: NGB-S--35--SE.

Huldén, M., B. Lund, G.P. Poulsen, E. Thörn, and J. Weibull (1998). The Nordic commitment: regional and international collaboration on plant genetic resources. Plant Varieties and Seeds 11: 1-13.

Hylander, N., Ö. Nilsson, I. Nordin, and H. Wanderoy (1948). *Kulturväxters namn på Svenska och Latin*. LTs Förlag, Borås, Sweden. ISBN 91-36-00281-X. [Second edition in 1960, third edition in 1977] [In Swedish]

IBPGR (1974). International board for plant genetic resources, first meeting, Rome, 5-7 June 1974. Consultative Group on International Agricultural Research (CGIAR), Rome, Italy. AGPE:IBPGR/74/6.

IBPGR (1976). First Report of the advisory committee on the genetic resources communication, information and documentation system (GR/CIDS). International Board for Plant Genetic Resources, Rome, Italy. AGPE:IBPGR/76/7.

IBPGR (1985) Ecogeographical surveying and *in situ* conservation of crop relatives, Report of an IBPGR task force, 30 July - 1 August, 1984, Washington, DC. IBPGR Secretariat, Rome, Italy. AGPG: IBPGR/84/132. 27 p.

IBPGR (1985). Cost-effective long-term seed stores. A report of the meeting of a sub-committee of the IBPGR advisory committee on seed storage. International Board for Plant Genetic Resources, Rome, Italy. 38 p. AGPG:IBPGR/85/217. [NordGen library: Qdda.4, id 492]

IBPGR (1990). Annual report 1989. International Board for Plant Genetic Resources (IBPGR), Rome, Italy. ISBN: 92-9043-197-0.

International Convention (1910). International Convention for the Creation of an International Agricultural Institute, Signed at Rome, June 7, 1905. [British Ratification deposited at Rome, 8 May 1907]. Treaty Series No 17. Available at http://www.fco.gov.uk/resources/en/pdf/treaties/TS1/1910/17, verified 4 Jan 2011.

International Convention (1930). Protocol regarding the International Convention of June 7, 1905, for the Creation of the International Institute of Agriculture, Rome, April 21, 1926. Treaty Series No 5. Available at http://www.fco.gov.uk/resources/en/pdf/treaties/TS1/1930/5, verified 4 Jan 2011.

IPCC (2007a). Climate change 2007: Synthesis report, summary for policymakers. Forth assessment report (AR4). Intergovernmental Panel on Climate Change (IPCC). Available at http://www.ipcc.ch/pdf/assessment-report/ar4/syr/ar4_syr_spm.pdf, verified 25 Jan 2011.

IPCC (2007b). Climate change 2007: The physical science basis. Contribution of working group I to the fourth assessment report of the Intergovernmental Panel on Climate Change [Solomon, S., D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M. Tignor, and H.L. Miller]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 996 p. Available from http://www.ipcc.ch/index.htm, verified 25 Jan 2011.

IPGRI (1992). Core collections: improving the management and use of plant germplasm collections. Proceedings of an IBPGR/CGN/ CENARGEN Workshop, 23 – 28 August 1992, Brasilia, Brazil. IPGRI, Rome, Italy.

IPGRI (1994). Genebank standards. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. ISBN: 92-9043236-5.

IPGRI (2001). European PGR information infra-structure. IPGRI Newsletter for Europe 22: 5.

IPGRI (2002). EPGRIS entering its third and final year. EURISCO and Central Crop Databases. IPGRI Newsletter for Europe 25: 5.

IPGRI (2003). Final EPGRIS Conference and ECP/GR Documentation and Information Network meeting. Newsletter for Europe 27: 4.

Iriondo, J.M., N. Maxted, and M.E. Dulloo (2008). Conserving plant genetic diversity in protected areas. CABI, Wallingford, Oxfordshire, UK. ISBN: 978-1-84593-282-4.

Jadav, V., L.M. Thorén, M. Veteläinen, and E. Thorn (2003). *Ex situ* preservation in permafrost as a complement to seed storage at the Nordic Gene Bank, p. 901–910. *In*: Smith, R.D., J.B. Dickie, S.H. Linington, H.W. Pritchard, and R.J. Probert (eds). Seed conservation: turning science into practice. Royal Botanic Gardens, Kew, UK. Available at http://www.kew.org/msbp/scitech/publications/SCTSIP_digital_book/pdfs/Chapter_50.pdf, verified 25 Jan 2011.

Jansky, S.H., R. Simon, and D.M. Spooner (2006). A test of taxonomic predictivity: Resistance to White mold in wild relatives of cultivated potato. Crop Science 46: 2561-2570. DOI: 10.2135/cropsci2005.12.0461.

Jansky, S.H., R. Simon, and D.M. Spooner (2008). A test of taxonomic predictivity to Early Blight in wild relatives of cultivated potato. Phytopathology 98(6): 680-687. DOI: 10.1094/PHYTO-98-6-0680.

Jansky, S.H., R. Simon, and D.M. Spooner (2009). A test of taxonomic predictivity: Resistance to the Colorado potato beetle in wild relatives of cultivated potato. Journal of Economic Entomology 102(1): 422-431.

Jarvis, A., M.E. Ferguson, D.E. Williams, L. Guarino, P.G. Jones, H.T. Stalker, J.F.M. Vallis, R.N. Pittman, C.E. Simpson, and P. Bramel (2003). Biogeography of wild *Arachis*: Assessing conservation status and setting future priorities. Crop Science 43(3): 1100-1108. DOI: 10.2135/cropsci2003.1100.

Jarvis, A., K. Williams, D. Williams, L. Guarino, P.J. Caballero, and G. Mottram (2005). Use of GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum* Sendtn.) in Paraguay. Genetic Resources and Crop Evolution 52: 671-682. DOI: 10.1007/s10722-003-6020-x.

Jarvis, A., A. Lane, and R.J. Hijmans (2008). The effect of climate change on crop wild relatives. Agriculture, Ecosystems and Environment 126: 13-23. DOI: 10.1016/j.agee.2008.01.013.

Jarvis, D.I., C. Padoch, and H.D. Cooper (eds) (2007). Managing biodiversity in agricultural ecosystems. Columbia University Press, New York, USA. 492 p. ISBN: 9780231136488.

Johannsen, W. (1903). *Om arvelighed i samfund og i rene linier*. Oversigt over det Kongelige Danske Videnskabernes Selskabs Forhandlinger, 3: 247-270. [In Danish] Available at http://caliban.mpiz-koeln.mpg.de/~stueber/johannsen/erblichkeit/index.html, verified 25 Jan 2011.

Johannsen, W.L. (1905) *Arvelighedslærens elementer* [The Elements of Heredity]. Copenhagen, Denmark. [In Danish] Available at http://caliban.mpiz-koeln.mpg.de/~stueber/johannsen/elemente/index.html, verified 25 Jan 2011.

Johnson, R.C. (2008). Gene banks pay big dividends to agriculture, the environment, and human welfare. PLoS Biology 6(6): e148. DOI: 10.1371/journal.pbio.0060148.

Johnson, R.C. and T. Hodgkin (eds) (1999). Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome, Italy. ISBN 92-9043-424-4.

Jones, P.G., S.E. Beebe, J. Tohme, and N.W. Galway (1997). The use of geographical information systems in biodiversity exploration and conservation. Biodiversity Conservation 6: 947-958.

Jones, P.G. and A. Gladkov (1999-2005). FloraMap: A computer tool for the distribution of plants and other organisms in the wild [Software]. CIAT, Cali, Columbia. [Last version is 1.03 from 2005]. ISSN: 958-694-078-0. Available from http://isa.ciat.cgiar.org/catalogo/producto.jsp?codigo=P328, verified 25 Jan 2011.

Jones, P.G., L. Guarino, and A. Jarvis (2002). Computer tools for spatial analysis of plant genetic resources data: 2. FloraMap. Plant Genetic Resources Newsletter 130: 1-6.

Jørgensen, J.H. and P. Kølster (1985). *Bygmeldugs udbredelse og betydning* [Powdery mildew distribution and impact]. Ugeskrift for Jordbrug 18: 459-463. [In Danish]

Joshi, S. and C.A. Viraktamath (2004). The sugarcane woolly aphid, *Ceratovacuna lanigera* Zehntner (Hemiptera: *Aphididae*): Its biology, pest status and control. Current Science 87(3): 307-316. Available at http://www.ias.ac.in/currsci/aug102004/307.pdf, verified 25 Jan 2011.

JPBI (2005). Genebank of the Jõgeva Plant Breeding Institute, Annual report 2005, http://www.sordiaretus.ee/files/GP/Aruanded_2005/JPBIeng.pdf, verified 25 Jan 2011.

Kampourakis, K. (2010). Mendel and the path to genetics: Portraying science as a social process. Science & Education [online first]. DOI: 10.1007/s11191-010-9323-2.

Kaur, K., K. Street, M. Mackay, N. Yahiaoui, and B. Keller (2008). Allele mining and sequence diversity at the wheat powdery mildew resistance locus *Pm3*. *In*: Appels, R., R. Eastwood, E. Lagudah, P. Langridge, M. Mackay, L. McIntyre, and P. Sharp (eds). The 11th International Wheat Genetics Symposium proceedings. Sydney University Press, Sydney, Australia. ISBN: 978-1-920899-14-1. Available at http://hdl.handle.net/2123/3227, verified 6 Jan 2011.

Kaur, N., K. Street, M. Mackay, N. Yahiaoui, and B. Keller (2008). Molecular approaches for characterization and use of natural disease resistance in wheat. European Journal of Plant Pathology 121(3): 387-397. DOI: 10.1007/s10658-007-9252-3. [Published July 2008]

Khoury, C., B. Laiberté, L. Guarino (2010). Trends in *ex situ* conservation of plant genetic resources: A review of global crop and regional conservation strategies. Genetic Resources and Crop Evolution 57: 625-639. DOI: 10.1007/s10722-010-9534-z.

Kiers, H.A.L. (2000). Towards a standard notation and terminology in multiway analysis. Journal of Chemometrics 14:105-122.

Kim, K.M. (1991) On the reception of Johannsen's Pure Line theory: Toward a sociology of scientific validity. Social Studies of Science 21(4): 649-679.

Kim, K.W., H.K. C.G.T. Cho, K.H. Ma, D. Chandrabalan, J.G. Gwag, T.S. Kim, E.G. Cho, and Y.J. Park (2007). PowerCore: a program applying the advanced M strategy with a heuristic search for establishing core sets. Bioinformatics 23 (16): 2155-2162. DOI: 10.1093/bioinformatics/btm313. [PowerCore is available for download from http://www.genebank.go.kr/eng/PowerCore/powercore.jsp, verified 25 Jan 2011]

Kloppenburg, J. and D.L. Kleinman (1987). The plant germplasm controversy. BioScience 37(3): 190-198. Stable URL: http://www.jstor.org/stable/1310518, verified 29 Dec 2010.

Kloppenburg, J.R. (1988). First the Seed: The political economy of plant biotechnology, 1492-2000. Cambridge University Press, Cambridge, UK. ISBN: 0-521-32691-5.

Knüpffer, H. (1988). The European barley database of the ECP/GR: An introduction. Kulturpflanze 36: 135-162.

Knüpffer, H. (1989). Identification of duplicates in the European barley database. p. 22-43. *In*: Report of a working group on barley (third meeting). European Cooperative Programme for the Conservation and Exchange of Crop Genetic Resources (EC/PGR). International Board for Plant Genetic Resources (IBPGR), Rome, Italy.

Knüpffer, H. (ed) (1999). *Index Seminum quae pro mutua commutatione offert Institut für Pflanzengenetik und Kulturpflanzenforschung Gatersleben 2000*. IPK, Gatersleben, Germany.

Knüpffer, H. and Th. van Hintum (2003). Summarized diversity - the Barley core collection. *In*: Bothmer, R. von, Th. van Hintum, H. Knüpffer, and K. Sato (eds). Diversity in barley (*Hordeum vulgare*), pp. 259–267. Elsevier Science B.V., Amsterdam, The Netherlands. ISBN: 9780444505859. DOI: 10.1016/S0168-7972(03)80015-4.

Knüpffer, H., J. Ochsmann, N. Biermann, and R. Narang (1998-2011). Mansfeld's world database of agricultural and horticultural plants [Online]. Leibniz Institute of Plant Genetics and Crop Plant Research (IPK Gatersleben), Gatersleben, Germany. Available at http://mansfeld.ipk-gatersleben.de/mansfeld, verified 25 Jan 2011.

Knüpffer, H., L.A. Morrison, A.A. Filatenko, K. Hammer, A. Morgounov, and I. Faberová (2002). English translation of the 1979 Russian taxonomic monograph of *Triticum* L. by Dorofeev et al.: project progress report. p. 49-55. *In*: Hernández, P., M.T. Moreno, J.I. Cubero, and A. Martin (eds). Triticeae IV. Proceedings from the 4th International Triticeae Symposium, September 10-12, 2001, Córdoba, Spain. Available at http://wheat.pw.usda.gov/ggpages/GrainTax/Manuscript_Dorofeev_extended.html, verified 25 Jan 2011.

Knüpffer, H., I. Terentyeva, K. Hammer, O. Kovaleva, and K. Sato (2003). Ecogeographic diversity - a Vavilovian approach. p. 53-76. *In*: Bothmer, R. von, T. van Hintum, H. Knüpffer, and K. Sato. Diversity in Barley (*Hordeum vulgare*). Developments in Plant Genetics and Breeding, 7. Elsevier Science B.V., Amsterdam, The Netherlands. 300 p. ISBN: 0-444-505857.

Knüpffer, H., J. Ochsmann, and N. Biermann (2003). The *"Mansfeld Database"* in its national and international context. p. 32-34. *In*: Knüpffer, H. and J. Ochsmann (eds). Rudolf Mansfeld and plant genetic resources. Proceedings of a symposium dedicated to the 100th birthday of Rudolf Mansfeld, Gatersleben, Germany, 8-9 October 2001. Zentralstelle für Agrardokumentation und -information (ZADI), Informationszentrum Biologische Vielfalt (IBV), Bonn, Germany. Available at http://www.genres.de/fileadmin/SITE_GENRES/downloads/schriftenreihe/Band22_Gesamt.pdf, verified 25 Jan 2011.

Knüpffer H., N. Biermann, D.T. Endresen, P. Kolasinski, W. Podyma, and J. de la Torre (2004). Genebanks as GBIF data providers the first experiences. *In*: Proceedings of TDWG 2004, Christchurch, New Zealand, October 10 to 17, 2004. Available at http://www.nhm.ac.uk/hosted_sites/tdwg/2004meet/paperabstracts/TDWG_2004_Papers_Knupffer_1.htm, verified 25 Jan 2011.

Knüpffer, H., D.T.F. Endresen, and S. Gaiji (2007). Integrating genebanks into biodiversity information networks. p. 34-35. *In*: 18th EUCARPIA Conference. Genetic Resources Section. Plant Genetic

Resources and their Exploitation in the Plant Breeding for Food and Agriculture. Piešťany, Slovak Republic. May 23 to May 26, 2007. ISBN: 9788088872634. Available at http://www.eucarpia.vurv.sk/abstracts/1/, http://www.eucarpia.org/03publications, verified 25 Jan 2011.

Kolodinska Brantestam, A. (2005). A century of breeding - is genetic erosion a reality? [PhD thesis]. Department of Crop Science, SLU. Acta Universitatis Agriculturae Sueciae. Vol. 2005:30.

Kolodinska Brantestam, A., R. von Bothmer, C. Dayteg, I. Rashal, S. Tuvesson, and J. Weibull (2004). Inter simple sequence repeat analysis of genetic diversity and relationships in cultivated barley of Nordic and Baltic origin. Hereditas 141: 186-192. DOI: 10.1111/j.1601-5223.2004.01867.x.

Komarov, V.L. (1934). *Flora SSSR, Tom I. Botanicheskiy Institut Akademii Nauk SSSR, Izdatel'stvo Akademii Nauk SSSR, Leningrad, USSR* [Flora of the USSR, Volume 1. The Botanical Institute of the Academy of Sciences of the USSR, Leningrad, USSR] [In Russian]

Komarov, V.L. (1968). Flora of the USSR, Volume I, *Archegoniatae* and *Embryophyta*. Smithsonian Institute, Washington, DC, USA, and The Israel Program for Scientific Translations (IPST) Press, Jerusalem, Israel. [Translated from Russian to English by N. Landau in 1968]. Available at http://www.archive.org/details/floraofussr01bota, verified 25 Jan 2011.

Konopka, J. and J. Hanson (eds) (1985). Information handling systems for genebank management. Workshop proceedings. International Board for Plant Genetic Resources (IBPGR), Rome, Italy / Nordic Gene Bank (NGB), Alnarp, Sweden.

Konopka, J., I. Kosareva, M. Mackay, O. Mitrofanova, K. Street, P. Strelchenko, J. Valkoun, E. Zuev, and M.F. Nawar (2005-2007) FIGS - Focused Identification of Germplasm Strategy [Online]. Bread Wheat Landrace Database. Available from http://figstraitmine.com/, verified 25 Jan 2011.

Konzak, C.F. and B. Sigurbjörnsson (1966). International cooperation in standardization of procedures in crop research data recording. Fifth Yugoslav symposium on research in wheat. Contemporary Agriculture 11-12: 691-696.

Koo, B., C. Nottenburg, and P.G. Pardey (2004). Plant and intellectual property: An international appraisal. Science 306(5700): 1295-1297. DOI: 10.1126/science.1106760.

Koo, B., P.G. Pardey, B.D. Wright, P. Bramel, D. Debrouck, M.E. van Dusen, M.T. Jackson, N.K. Rao, B. Skovmand, S. Taba, and J. Valkoun (2004). Saving seeds: The economics of conserving crop genetic resources *ex situ* in the future harvest centres of the CGIAR. CABI Publishing, Wallingford, Oxfordshire, UK. 232 p. ISBN-10: 0-85199-859-3.

Kumarasinghe, N.C. and B.R.S.B. Basnayake (2009). Influence of monsoonal weather on sudden establishment of the sugarcane woolly aphid in Sri Lanka. Sugar Tech 11(3): 267-273. DOI: 10.1007/s12355-009-0046-0.

Kurki, A. (ed) (1986). Apricot (*Prunus armenica*) from ECP/GR Prunus Central Data Base, July 1986. European Cooperative Programme for the conservation and exchange of crop Genetic Resources (ECP/GR) and Nordic Gene Bank (NGB), Alnarp, Sweden.

Latha R., L. Rubia, J. Bennett, and M.S. Swaminathan (2004). Allele mining for stress tolerance genes in *Oryza* species and related germplasm. Molecular Biotechnology 27(2): 101-108. DOI: 10.1385/MB:27:2:101.

Lawrence, T. (ed) (1984). Collection of crop germplasm, the first ten years, 1974-84. International Board for Plant Genetic Resources (IBPGR), Rome, Italy. AGPG:IBPGR/84/106.

Lehmann, C.O. (1981). Collecting European land-races and development of European gene banks - historical remarks. Genetic Resources and Crop Evolution 29(1): 29-40. DOI: 10.1007/BF02014732.

Lindeberg, G., K. Sandvad, and S. Blixt (1981). *PM rörande Nordiska genbankens databehandlingsstruktur och dess utbygnad samt uppdateringsbehov av nuvarande ADB-anläggning*. Nordiska Genbanken, Lund. (Retrieved from the NordGen correspondence archive, cornum 8100154 and cornum 8100030). [In Swedish]

Linnaeus, C. (1737). *Critica Botanica* [Critique of botany]. Leiden, Germany. Available at http://books.google.com/books?id=JHsZAAAAYAAJ, verified 25 Jan 2011.

Linnaeus, C. (1753). *Species Plantarum. Laurentii Salvii, Holmiae* [Stockholm], Sweden. 1200 p. Available at http://www.biodiversitylibrary.org/bibliography/669, http://www.botanicus.org/item/31753000802824, verified 25 Jan 2011.

Linnestad, C. (2001). *Fra arvelære til bioteknologi*. Genialt 10(2): 6. Available at http://www.bion.no/filarkiv/2010/07/genialt2-2001.pdf, verified 25 Jan 2011.

Lipman, E., M.W.M. Jongen, Th.J.L. van Hintum, T. Grass, and L. Maggioni (eds) (1997). Central crop databases: Tools for plant genetic resources management. International Plant Genetic Resources Institute, Rome, Italy; and CGN, Wageningen, Netherlands. ISBN 92-9043-320-5.

Lira Saade, R. (1996). *Estudios taxonomics y ecogeograficos de las cucurbitaceae latinoamericanas de importancia economica*. Systematic and ecogeographic studies on crop genepools 9. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. 281 p. ISBN: 92-9043-263-2. [In Spanish]

Liu, K.J. and Muse, S.V. (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21: 2128-2129. DOI: 10.1093/bioinformatics/bti282.

Liu, X., J.L. Marshall, P. Stary, O. Edwards, G. Puterka, L. Dolatti, M. EL Bouhssini, J. Malinga, J. Lage, and C.M. Smith (2010). Global phylogenetics of *Diuraphis noxia* (Hemiptera: *Aphididae*), an invasive aphid species: Evidence for multiple invasions into North America. J. Econ. Entomol. 103(3): 958-965. DOI: 10.1603/EC09376.

Löfgren, A. (2009). *Kostnader 2008 och estimering av framtida kostnader för genbanken vid NordGen Växter. Genomförd på Uppdrag av direktör och styrelse för NordGen 28 maj 2009* [Operation costs 2008 and estimated future costs for the gene bank at NordGen Plants. Requested by the director and the board of NordGen 28 May 2009]. Ernst & Young, Malmö, Sweden. [In Swedish]

Loskutov, I.G. (1999). Vavilov and his Institute: a history of the world collection of plant genetic resources in Russia. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. 188 p. ISBN: 978-92-9043-412-2.

Lyman, J.M. (1984). Progress and planning for germplasm conservation of major food crops. Plant Genetic Resources Newsletter 60: 3-21.

Lysenko, T.D. (1951). Heredity and its variability. University Press of the Pacific, Honolulu, Hawaii. ISBN: 0-89875-660-X. [Reprinted in 2001 from the 1951 edition] Available at http://books.google.com/books?hl=en&lr=&id=A4KSlbYtIe0C, verified 25 Jan 2011.

Mackay, M.C. (1986). Utilizing wheat genetic resources in Australia. p. 56-61. *In*: McLean, R. (ed). Proceedings of the 5th assembly wheat breeders' society in Australia, Merredin 18-22 Aug 1986. Western Australian Department of Agriculture, Perth, Australia. 580 p. ISBN: 9780730913269.

Mackay, M.C. (1990). Strategic planning for effective evaluation of plant germplasm. p. 21–25. *In*: Srivastava J.P., and A.B. Damania (eds). Wheat genetic resources: Meeting diverse needs. John Wiley & Sons, Chichester, UK. ISBN 0-471-92880-1.

Mackay, M.C. (1995). One core collection or many? p. 199-210. *In:* Hodgkin T., A.H.D. Brown, Th.J.L. van Hintum, and A.A.V. Morales (eds). Core collections of plant genetic resources. Proceedings from the IBPGR/CGN/CENARGEN workshop on 'Core Collections: Improving the Management and Use of Plant Germplasm Collections', held in Brasilia August 1992. John Wiley & Sons, Chichester, UK. 269 p. ISBN: 978-0-471-95545-0.

Mackay M.C. and K. Street (2004). Focused identification of germplasm strategy – FIGS. p. 138-141. *In*: Black, C.K., J.F. Panozzo, and G.J. Rebetzke (eds). Cereals 2004. Proceedings of the 54th Australian Cereal Chemistry Conference and the 11th Wheat Breeders' Assembly, 21-24 September 2004, Canberra, Australian Capital Territory (ACT). Cereal Chemistry Division, Royal Australian Chemical Institute, Melbourne, Australia. URL: http://books.google.com/books?id=SlYbNAAACAAJ, verified 25 Jan 2011.

Maggioni, L. (ed) (2005). Summary of a network coordinating group on documentation and information and the EURISCO advisory group. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. Available at http://www.bioversityinternational.org/fileadmin/bioversity/publications/pdfs/1051.pdf, verified 25 Jan 2011.

Maggioni, L. (ed) (2007). Minutes of a joint meeting of the documentation and information network coordinating group and the EURISCO advisory group. Planning for the continuation of EPGRIS, 2-3 April 2007. Bioversity International, Rome, Italy. Available at http://www.ecpgr.cgiar.org/Networks/Info_doc/DINCG_AdvGR_ROMEApril07.pdf, verified 25 Jan 2011.

Maggioni, L. (ed) (2010). Report of the ECPGR documentation and information network coordinating group, forth meeting, 17-18 February 2010, Maccarese, Rome, Italy. Bioversity International, Rome, Italy. 22 p. Available at http://www.ecpgr.cgiar.org/Networks/Info_doc/Doc&Info NCG Fourth Meeting final for Web 260510.pdf, verified November 4, 2010.

Maggioni, L., P. Marum, N.R. Sackville Hamilton, M. Huldén, and E. Lipman (compilers) (2000). Report of a working group on forages. Seventh meeting, 18-20 November 1999, Elvas, Portugal. European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR), International Plant Genetic Resources Institute (IPGRI), Rome, Italy. ISBN: 92-9043-451-1.

Malthus, T.R. (1798). An essay on the principle of population, as it affects the future improvement of society. J. Johnsen, London, UK. Available at http://en.wikisource.org/wiki/An_Essay_on_the_Principle_of_Population, verified 25 Jan 2011.

Maltzev A.I. (1914). *Iz nablyudeniy nad razvitiem dikorastushchikh i sornykh ovsov* [On the development of wild and weedy oats]. Bulletin of the Bureau of Applied Botany 7(12): 786-791. [In Russian]

Mansfeld, R. (1959). *Vorläufiges Verzeichnis landwirtschaftlich oder gärtnerisch kultivierter Pflanzenarten (mit Ausschluß von Zierpflanzen).* Kulturpflanze, Beiheft 2. 659 p. [In German] [approximately 1430 cultivated plant species; updated second issue: Schultze-Motel, 1986; 3rd edition translated from German to English: Hanelt and IPK, 2001]

Mansfeld, R. (1962). *Über "alte" und "neue" Systematik der Pflanzen*. Kulturpflanze, Beiheft 3: 26-46.

Maredia, M.K., R. Bernsten, and C. Ragasa (2010). Returns to public sector plant breeding in the presence of spill-ins and private goods: the case of bean research in Michigan. Agricultural Economics 41(5): 425-442. DOI: 10.1111/j.1574-0862.2010.00455.x.

Marshall, D.R. and A.H.D. Brown. (1975). Optimum sampling strategies in genetic conservation. p. 53–80. *In*: Frankel, O.H., and J.G. Hawkes (eds). International Biological Programme, volume 2. Crop genetic resources for today and tomorrow. Cambridge Univ. Press, Cambridge, UK. ISBN: 0-521-20575-1.

Maxted, N. (1995). An ecogeographical study of *Vicia* subgenus *Vicia*. Systematic and ecogeographic studies on crop genepools 8. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. 184 p. ISBN: 92-9043-240-3.

Maxted, N., B.V. Ford-Lloyd, and J.G. Hawkes (eds) (1997). Plant genetic conservation: The *in situ* approach. Chapman & Hall, London, UK. ISBN: 0-412-63730-8.

Maxted, N., P. Mabuza-Diamini, H. Moss, S. Padulosi, A. Jarvis, and L. Guarino (2004). An ecogeographic study of African *Vigna.* Systematic and ecogeographic studies on crop genepools 11. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. 454 p. ISBN: 92-9043-637-9.

Maxted N., S.P. Kell, B.V. Ford-Lloyd, E. Dulloo, and J. Iriondo (2008). Crop wild relative conservation and use. CABI, Wallingford, Oxfordshire, UK. ISBN: 978-1-84593-099-8.

McCouch, S. (2004). Diversifying selection in plant breeding. PLoS Biology 2(10): 1507-1512. DOI: 10.1371./journal.pbio.0020347.

Meganck, B., D. Meirte, P. Mergen, and F. Theeten (2006). Providing itinerary related datasets and tools for integration, visualisation and quality check - system specifications. Milestone report for SYNTHESYS network activity NA-D: Developing and maintaining databases. Royal Museum for Central Africa (RMCA), Tervuren, Belgium. Available from http://www.biocase.org/products/geo_services/itineraries/files/SYNTHESYS_milestone_report_may31.pdf, verified 25 Jan 2011.

Mendel, G. (1866). *Versuche uber Pflanzen-Hybriden* [Experiments on Plant Hybrids]. Verhandlungen des naturforschenden Vereines in Brunn 4: 3-47. [Translated to English by William Bateson in 1901,

updated by Roger Blumberg in 1995]. Available at http://www.mendelweb.org/Mendel.html and at http://www.esp.org/foundations/genetics/classical/gm-65.pdf, verified 25 Jan 2011.

Mitrofanova, O.P., P.P. Strelchenko, A.V. Konarev, and F. Balfourier (2009). Genetic differentiation of hexaploid wheat inferred from analysis of microsatellite loci. Russian Journal of Genetics 45(11): 1351-1359. DOI: 10.1134/S102279540911009X.

Monaghan, F.V. and A.F. Corcos (1986). Tschermak: a non-discoverer of Mendel's: I. An historical note. Journal of Heredity 77(6): 468-469.

Monaghan, F.V. and A.F. Corcos (1987). Tschermak: a non-discoverer of Mendelism II. A critique. Journal of Heredity 78(3): 208-210.

Moore, J.D., S.P. Kell, J.M. Iriondo, B.V. Ford-Loyd, and N. Maxted (2008). CWRML: representing crop wild relative conservation and use data in XML. BMC Bioinformatics 9: 116. DOI: 10.1186/1471-2105-9-116. Schema available at http://pgrforum.org/CWRML.htm, verified 25 Jan 2011.

Morrison, L.A., I. Faberová, A.A. Filatenko, K. Hammer, H. Knüpffer, A. Morgounov, and S. Rajaram (2000). Call to support an English translation of the 1979 Russian taxonomic monograph of *Triticum* by Dorofeev *et al.* Wheat Information Service 90: 52-53. Available at http://wheat.pw.usda.gov/ggpages/GrainTax/Wisdorofeev2.html, verified 25 Jan 2011.

Motley, T.J., N. Zerega, and H. Cross (eds) (2006). Darwin's harvests, new approaches to the origins, evolution, and conservation of crops. Columbia University Press, New York, USA. ISBN: 0-231-13316-2.

Mukherjee, S.K. (1985). Systematic and ecogeographic studies on crop genepools: 1. *Mangifera* L. IBPGR, Rome, Italy. 86 p.

Nabhan, G.P. (1990). Wild *Phaseolus* ecogeography in the Sierra Madre Occidental, Mexico: Areographic techniques for targeting and conserving species diversity. Systematic and ecogeographic studies on crop genepools 5. IBPGR, Rome, Italy. 35 p. ISBN 92-9043-198-9.

Nabhan, G.P. (2009). Where our food comes from: retracing Nikolay Vavilov's quest to end famine. Island Press, Washington DC, USA. ISBN: 978-1-59726-399-3.

Nägeli, C.W. von (1884). *Mechanisch-physiologische Theorie der Abstammungslehre* [A mechanico-physiological theory of organic evolution]. Drück und Verlag von R. Oldenburg, München and Leipzig, Germany. Available from http://vlp.mpiwg-berlin.mpg.de/library/data/lit19841/index_html verified 25 Jan 2011. [In German, See Nägeli, 1898 for the English version]

Nägeli, C.W. von (1898). A mechanico-physiological theory of organic evolution. [Second edition]. The Open Court Publishing Co., Chicago, USA. Available from http://www.gutenberg.org/ebooks/33514, verified 25 Jan 2011.

Negri, V., N. Maxted, and M. Veteläinen (2009). European landrace conservation: An introduction. p. 1-22. *In*: Veteläinen M., V. Negri, and N. Maxted (eds). European landraces on-farm conservation, management and use. Bioversity Technical Bulletin No. 15. Bioversity International, Rome, Italy. 344 p. ISBN: 978-92-9043-805-2.

NGB (1987). Requirements to material of and data on self-pollinating cereal species (*Avena*, *Hordeum*, and *Triticale*/*Secalotricum*) to be stored at the Nordic Gene Bank. Nordic Gene Bank (NGB), Alnarp, Sweden. [Compiled by the NGB working group on cereal species, March 3, 1987. J. Helms Jørgensen (chairman). Approved by the board of NGB, 20 March 1987]

NGB (1989). Seed catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-06-4.

NGB (1990). Requirements to material of and data on cross-pollinating cereal species to be stored in the Nordic Gene Bank. Nordic Gene Bank (NGB), Alnarp, Sweden. [Compiled at the NGB working group on cereal species, at Risø in Denmark, 8 October 1990]

NGB (1991). Nordiska Genebanken 1979 - 1989. Nordic Gene Bank, Alnarp, Sweden. ISSN: 1100-3456. [10-year jubilee, with summary from the first 10 years of operation]

NGB (1994). Strategi inför år 2000 [Strategy before the year 2000]. Nordic Gene Bank (NGB), Alnarp, Sweden. ISSN: 1100-3456. ISRN: NGB-S--26--SE. [NGB Publications No. 26] [In Swedish]

NGB (2002a). Avtal - Regis ekonomisk förening [Contract with Regis on georeferencing of genebank accessions]. Nordic Gene Bank archive dated 14 October 2002. [In Swedish]

NGB (2002b). Internal report on a georeferencing project for accessions in SESTO. [In Swedish] Description of the collecting missions conducted by the Nordic Gene Bank is available from http://sesto.nordgen.org/sesto/index.php?scp=ngb&thm=projects&mod=prolst&prowgr=PRR&progrpide=14, verified 25 Jan 2011.

Nilsson-Ehle, H. (1911). *Viktigare framsteg under de senare åren med afseende på de teoretiska grundvalarna för växtförädlingen. Mendelismen och dess betydelse*. Aktiebolaget Södermanlands Läns Tidings Tryckeri, Nyköping, Sweden. 24 p. Available at http://vlp.mpiwg-berlin.mpg.de/library/data/lit29282, verified 25 Jan 2011. [In Swedish]

Nilsson-Ehle, Herman (1919). *Något om ärftlighets-vetenskapens praktiska och ekonomiska betydelse*. Landskrona Tryckeri, Landskrona, Sweden. Available at http://vlp.mpiwg-berlin.mpg.de/library/data/lit29293, verified 25 Jan 2011. [In Swedish].

Nilsson, A. and R. von Bothmer (2010). Measures to promote Nordic plant breeding. TemaNord 2010:518. Nordic Council of Ministers, Copenhagen, Denmark. ISBN: 978-92-893-2000-9. Available online at http://www.norden.org/en/publications/publications/2010-518, verified 25 Jan 2011.

Ninnes, P., T. Payne, and J. Lawrence (2002). SINGER is music to crop scientists. Partners in Research for Development 15: 41-44. Available at http://www.sgrp.cgiar.org/?q=node/659, verified 6 Feb 2011.

Nobel Foundation (1970). Norman Borlaug - biography [Online]. Available at http://nobelprize.org/nobel_prizes/peace/laureates/1970/borlaug-bio.html, verified 15 Jan 2011. [From Nobel Lectures, Peace 1951-1970, Editor Frederick W. Haberman, Elsevier Publishing Company, Amsterdam, 1972]

NordGen (2009). Cereal pre-breeding workshop [Online]. Available at http://www.nordgen.org/index.php/skand/Innehaall/Aktiviteter/Cereal-Pre-Breeding-Workshop, verified 25 Jan 2011.

Palmer, R.G. and J.J. Doyle (2009). Dedication: Anthony H. D. Brown conservation geneticist. p. 1-21. *In*: Janick, J. (ed). Plant breeding reviews, volume 31. John Wiley & Sons, Inc., Hoboken, NJ, USA. DOI: 10.1002/9780470593783.ch1. ISBN: 9780470593783.

Palmova, E.F. (1935). *Ogiz Selkhozgiz* [Introduction in wheats ecology]. Leningrad and Moscow. [In Russian] [*cf.* Strelchenko et al., 2004]

Palmstierna, H., G. Julen, L. Kåhre, and S.O. Myresten (1975). *Genbank för jordbruk och trädgårdsnäring: Betänkande avgivet av Genbankutredningen* [Genebank for agriculture and horticulture: Report from the Genebank committee]. Jordbruksdepartementet, Liber Förlag, Stockholm, Sweden. 78 p. ISSN: 0346-5667, ISBN: 9138023539. Ds Jo 1975:5. [In Swedish with English summary]

Peeters, J.P. and J.A. Martinelli (1989). Hierarchical cluster analysis as a tool to manage variation in germplasm collection. Theoretical and Applied Genetics 78: 42-48.

Peeters, J.P., and J.T. Williams (1984). Towards better use of genebanks with special reference to information. Plant Genetic Newsletter 60: 22-32.

Peeters, J.P., H.G. Wilkes, and N.W. Galway (1990). The use of ecogeographical data in the exploitation of variation from gene banks. Theoretical and Applied Genetics 80: 110-112.

Pessoa-Filho, A., P.H.N. Rangel, and M.E. Ferreira (2010). Extracting samples of high diversity from thematic collections of large gene banks using a genetic-distance based approach. BMC Plant Biology 10: 127. DOI: 10.1186/1471-2229-10-127.

Pistorius, R. (1997). Scientists, plants and politics: a history of the plant genetic resources movement. International Plant Genetic Resources Institute, Rome. 134 p. ISBN: 9789290433088. [Not cited in the text]

Pistrick, K. (2003). Mansfeld's encyclopedia of agricultural and horticultural crops and the Mansfeld phenomenon. p. 21-31. *In*: Knüpffer, H. and J. Ochsmann (eds). Rudolf Mansfeld and plant genetic resources. Proceedings of a symposium dedicated to the 100th birthday of Rudolf Mansfeld,

Gatersleben, Germany, 8-9 October 2001. Zentralstelle für Agrardokumentation und -information (ZADI), Informationszentrum Biologische Vielfalt (IBV), Bonn, Germany. Available at http://www.genres.de/fileadmin/SITE_GENRES/downloads/schriftenreihe/Band22_Gesamt.pdf, verified 25 Jan 2011.

PlantExplorers.com (1999-2010). Plant explorers - the adventure is growing [Online]. Available at http://plant explorers.com, verified 25 Jan 2011.

Plucknett, D.L., N.J.H. Smith, J.T. Williams, and N.M. Anishetty (1987). Gene banks and the world's food. Princeton University Press, Princeton, NJ, USA. ISBN: 0-691-08438-6. Available at http://books.irri.org/getpdf.htm?book=0691084386, verified 25 Jan 2011.

Prada, D. (2009). Molecular population genetics and agronomic alleles in seed banks: Search for a needle in the haystack? Journal of Experimental Botany 60: 2541-2552. DOI: 10.1093/jxb/erp130.

Pretorius, Z.A., R.P. Singh, W.W. Wagoire, and T.S. Payne (2000). Detection of virulence to wheat stem rust resistance gene *Sr31* in *Puccinia graminis* f. sp. *tritici* in Uganda. Plant Disease 84(2): 203. DOI: 10.1094/PDIS.2000.84.2.203B.

Pringle, P. (2008). The murder of Nikolai Vavilov: The story of Stalin's persecution of one of the greatest scientists of the twentieth century. Simon & Schuster, New York, USA. ISBN-13: 978-0-7432-6498-3.

Proskowetz, E. von (1890). *Welches Werthverhältnis besteht zwischen den Landrassen landwirtschaftlicher Culturpflanzen und den sogenannten Züchtungsrassen?* [What relationship exists between the values of landraces and the breeding lines?] Internationaler land- und forstwirtschaftlicher Congress zu Wien 1890. Section I: Landwirtschaft. Subsection Pflanzenbau, Frage 5. Heft 13: 3-18.

QGIS (2002-2011). Quantum GIS [Software]. Available for download from http://qgis.org/, verified 25 Jan 2011.

Qualset, C.O. and H.L. Sands (2005). Safeguarding the future of U.S. agriculture: The need to conserve threatened collections of crop diversity worldwide. University of California, Division of Agriculture and Natural Resources Conservation Program. Davis, CA, USA. ISBN: 92-9043-671-9.

Qvenild, M. (2005). Securing seeds in permafrost: an idea whose time has come [Master thesis, supervisor Cary Fowler]. NorAgric, Norwegian University of Life Sciences, Ås, Norway. 81 p. Available at http://www.umb.no/statisk/noragric/publications/master/2006_marte_qvenild.pdf, verified 25 Jan 2011.

Qvenild, M. (2006). Svalbard Global Seed Vault: a 'Noah's Ark' for the world's seeds. Development in Practice 18(1): 110-116. DOI: 10.1080/09614520701778934.

Ramanna, A. (2006). Farmers' rights in India: A Case Study. The Fridtjof Nansens Institute, Lysaker, Norway. ISBN: 82-7613-491-2.

Ramírez-Villegas J., C. Khoury, A. Jarvis, D.G. Debouck, and L. Guarino (2010). A gap analysis methodology for collecting crop genepools: A case study with *Phaseolus* beans. PLoS ONE 5(10): e13497. DOI: 10.1371/journal.pone.0013497.

Rana, R.S., R.L. Sapra, R.C. Agrawal, and R. Gambhir (1991). Plant genetic resources documentation and information management. National Bureau of Plant Genetic Resources (NBPGR), New Delhi, India.

Rao, N.K., J. Hanson, M.E. Dulloo, K. Ghosh, D. Nowell, and M. Larinde (2006). Manual of seed handling in genebanks. Handbooks for genebanks no. 8. Bioversity International, Rome, Italy. 147 p. ISBN: 9789290437406.

Roberts, H.F. (1929). Plant Hybridization before Mendel. Princeton University Press, Princeton, New Jersey, USA. Available at http://www.archive.org/details/planthybridizati00robe, verified 30 Dec 2010.

Roll-Hansen, N. (2009). Sources of Wilhelm Johannsen's genotype theory. Journal of the History of Biology 42(3): 457-493. DOI: 10.1007/s10739-008-9166-8.

Rubenstein, K.D., P. Heisey, R. Shoemaker, J. Sullivan, and G. Frisvold (2005). Crop genetic resources, an economic appraisal. Economic Information Bulletin, 2. Economic Research Service, USDA. 47 p. Available at http://www.ers.usda.gov/publications/eib2/, verified 25 Jan 2011.

Ruttan, V.W. (2002). Productivity growth in world agriculture: Sources and constraints. Journal of Economic Perspectives 16(4): 161-184. DOI: 10.1257/089533002320951028. Stable URL: http://www.jstor.org/stable/3216919, verified 25 Jan 2011.

Rydström, G. (1989). BIRS, biological information retrieval system user's guide [Software]. Nordic Gene Bank, Alnarp, Sweden.

Sánchez G., J.J. and L. Ordaz S. (1987). *El Teocintle en México: Distribución actual de poblaciones.* Systematic and ecogeographic studies on crop genepools: 2. IBPGR, Rome, Italy. 50 p.

Schischkin, B.K. and E.G. Bobrov (eds) (1960). *Flora SSSR, Tom 30. Botanicheskii institut im. V.L. Komarova, Akademii Nauk SSSR, Moscow-Leningrad, SSSR.* [Flora of the USSR, Volume 30. V.L. Komarov Botanical Institute, Academy of Sciences of the USSR, Moscow-Leningrad, USSR].

Schischkin, B.K. and E.G. Bobrov (eds) (2002). Flora of the USSR, Volume 30, Compositae, Genus *Hieracium*. Smithsonian Institute, Washington, DC, USA, and Amerind Publishing Co Pvt Ltd, New Delhi, India. [Translated from Russian to English by Dr. V.S. Dhote; scientific editors were S.W. Shetler, G.N. Fet, and E.A. Unumb]. Available at http://www.archive.org/details/floraofussr30bota, verified 25 Jan 2011.

Schoen D.J. and A.H.D. Brown (1993). Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 22: 10623–10627.

Schoen, D.J. and A.H.D. Brown (1995). Maximizing genetic diversity in core collections of wild relatives of crop species. p. 55-76. *In*: Hodgkin, T., A.H.D. Brown, T.J.L. van Hintum, and E.A.V. Morales (eds). Core collections of plant genetic resources. John Wiley & Sons, UK. 269 p. ISBN: 978-0-471-95545-0.

Schultze-Motel, J. (ed) (1986). *Rudolf Mansfelds Verzeichnis landwirtschaftlicher und gärtnerischer Kulturpflanzen (ohne Zierpflanzen)*. 4 volumes, 2013 p. Akademie-Verlag, Berlin, Germany. [In German] [Approximately 4800 cultivated plant species].

Serwiński, J., and J. Konopka (1984). European catalogue of genus *Secale* L. First edition. European Cooperative Programme for the Conservation and Exchange of Crop Genetic Resources, International Board for Plant Genetic Resources, Rome, Italy.

SESTO (2002-2011). SESTO genebank management system [Online database]. Available at http://sesto.nordgen.org, verified 25 Jan 2011.

Settar, N. (1982). *Türki Bitki Genetik Kaynaklari* [Plant genetic resources of Turkey] Index Seminum (1970-1980). Ege Bölge Zirai Arastirma Enstitüsü Yayinlari No 22. Menemen-Izmir, Turkey.

SGSV Portal (2008-2011). Svalbard Global Seed Vault seed portal [Online database]. Nordic Genetic Resource Center (NordGen), Alnarp, Sweden. Available at http://www.nordgen.org/sgsv, verified 25 Jan 2011.

Shetler, S.G. (1967). The Komarov Botanical Institute, 250 years of Russian research. Smithsonian Institution Press, Washington, D.C., USA. Available at http://www.archive.org/details/komarovbotanical00shet, verified 25 Jan 2011.

Simmonds, N.W. (ed) (1976). Evolution of crop plants. Longman, London, UK. 339 p. ISBN: 0-582-44496-9.

Singh, P.R., D.P. Hodson, J. Huerta-Espino, Y. Jin, P. Njau, R. Wanyera, S.A. Herrera-Foessel, and R.W. Ward (2008). Will stem rust destroy the world's wheat crop? Advances in agronomy, 98: 271-309. DOI: 10.1016/S0065-2113(08)00205-8.

Skofic M., E. van Strien, D.T. Endresen, and E. Arnaud. (2009). Registering and uploading datasets in the Generation CP Central Registry (SP4), 2009 GCP Annual Research Meeting, 20-23 September 2009, Bamako, Mali.

Skovmand, B. (1973). Variation in wheat characteristics involved in morphological resistance to wheat stem rust [Master thesis]. University of Minnesota, Minneapolis, MN, USA. 41 p. Information

available at http://www.worldcat.org/title/variation-in-wheat-characteristics-involved-in-morphological-resistance-to-wheat-stem-rust/oclc/62623757&referer=brief_results, verified 25 Jan 2011.

Skovmand, B. (1976). Inheritance of slow rust development in spring wheat and the relationship of slow rusting to specific resistance [PhD thesis]. University of Minnesota, Minneapolis, MN, USA. 114 p. Information available at http://www.worldcat.org/title/inheritance-of-slow-rust-development-in-spring-wheat-and-the-relationship-of-slow-rusting-to-specific-resistance/oclc/62480632&referer=brief_results, verified 25 Jan 2011.

Skovmand, B., M.P. Reynolds, and I.H. DeLacy (2001). Mining wheat germplasm collections for yield enhancing traits. Euphytica 119: 25-32.

Skovmand, B., M.C. Mackay, and K. Street (2004). A new approach to locating and utilizing oat genetic resources. p. 33-37. *In*: Peltonen-Sainio, P., and M. Topi-Hulmi (eds). Proceedings 7th International Oat Conference, Agrifood Research Reports 51. MTT Agrifood Research Finland, Jokioinen, Finland. 239 p. ISBN: 951-729-880. Available at http://www.mtt.fi/met/pdf/met51.pdf, verified 6 Jan 2011.

Smale, M. (1997). The green revolution and wheat genetic diversity: some unfounded assumptions. World Development 25: 1257-1269.

Smale, M., M.P. Reynolds, M. Warburton, B. Skovmand, R. Trethowan, R.P. Singh, I. Ortiz-Monasterio, J. Crossa, M. Khairallah, and M. Almanza (2001). Dimensions of diversity in CIMMYT bread wheat from 1965 to 2000. CIMMYT, Mexico. Available at http://apps.cimmyt.org/Research/Wheat/map/research_results/DimDiversity/DimDiversity_contents.htm, verified 25 Jan 2011.

Smartt, J. and N.W. Simmonds (eds) (1995). Evolution of crop plants. Second edition. Longman Scientific & Technical, Harlow, Essex, UK. 531 p. ISBN: 0-582-08643-4. [First published in 1976]

Smilde, A., R. Bro, and P. Geladi (2004). Multi-way analysis, applications in the chemical sciences. Wiley, Chichester, UK. ISBN: 978-0-471-98691-1.

Smith, R.D. (1989). Technical Evaluation of Permafrost Storage in Svalbard. Kew Gardens, London, UK and International Board for Plant Genetic Resources (IBPGR), Rome, Italy.

Smith, R.D., J.B. Dickie, S.H. Linington, H.W. Pritchard, and R.J. Probert (eds) (2003). Seed conservation: turning science into practice. Royal Botanic Gardens Kew, Kew, UK. 1023 p. Available at http://www.kew.org/msbp/scitech/publications/sctsip.htm, verified 25 Jan 2011.

SNSK and NGB (2004). *Kontrakt mellom Store Norske Spitsbergen Kullkompani AS and Nordisk Genbank for landbruks- og havebrugsplanter*. Signed by SNSK Longyearbyen 13/7-84, and NGB Lund 30/7-84. [NordGen archives] [In Scandinavian]

Spagnoletti Zeuli, P.L. and C.O. Qualset (1993). Evaluation of five strategies for obtaining a core subset from a large genetic resource collection of durum wheat. Theoretical and Applied Genetics 87: 295-304. DOI: 10.1007/BF01184915.

Spooner, D.M., S.H. Jansky, and R. Simon (2009). Test of taxonomic and biogeographic predictivity: Resistance to disease and insect pests in wild relatives of cultivated potato. Crop Science 49: 1367-1376. DOI: 10.2135/cropsci2008.04.0211.

Srivastava J.P. and A.B. Damania (eds) (1990). Wheat genetic resources: Meeting diverse needs. John Wiley & Sons, Chichester, UK. ISBN 0-471-92880-1.

Stafleu, F.A. (1969). Botanical gardens before 1818. Boissiera 14: 13-46.

Stansfield, W.D. (2006). Luther Burbank: Honary member of the American Breeders' Association. Journal of Heredity 97(2): 95-99. DOI: 10.1093/jhered/esj015.

Statsbygg (2008). Svalbard Global Seed Vault, Longyearbyen, Svalbard, new construction. Ferdigmelding nr 671/2008, Project no 11098. Statsbygg, Oslo, Norway. 28 p. Available at http://www.statsbygg.no/FilSystem/files/ferdigmeldinger/671_svalbard_frohvelv.pdf, verified 18 Jan 2011.

Stearn, W.T. (1986). Historical survey of the naming of cultivated plants. Acta Horticulturae 182: 19-28.

Stearn, W.T. (ed) (1953). International code of nomenclature for cultivated plants. Royal Horticultural Society, London, UK.

Stevens, S. S. (1946). On the theory of scales of measurement. Science 103(2684): 677–680. DOI: 10.1126/science.103.2684.677. Available at http://www.sciencemag.org/cgi/rapidpdf/103/2684/677, verified 25 Jan 2011.

Stockwell, D. (2007). Niche modeling: Predictions from statistical distributions. Chapman and Hall/CRC. ISBN: 9781584884941.

Street, K. and M. Mackay (2004). Revolutionizing plant genetic resources management, trait discovery & utilization: A project concept note. ICARDA, Aleppo, Syria, and Australian Winter Cereal Collection (AWCC), Tamworth, Australia. [Project Concept Note].

Street, K., E. De Pauw, J. Ryan, and M.C. Mackay (2004). Focused Identification of Germplasm Strategy: Identifying wheat landraces for salinity screening in Eurasia. *In*: Proceeding from ASA CSSA SSSA International Annual Meetings, 31 October to 4 November 2004, Seattle, Washington, USA. Available at http://download.clib.psu.ac.th/datawebclib/e_resource/e_database/agronomy/2004/Browse/pdf/ACS/3604.pdf, verified 6 Jan 2011.

Street, K., M. Mackay, E. Zuev, N. Kaul, M. El Bouhssini, J. Konopka, and O. Mitrofanova (2008). Swimming in the genepool - a rational approach to exploiting large genetic resource collections. *In*: Appels, R., R. Eastwood, E. Lagudah, P. Langridge, M. Mackay, L. McIntyre, and P. Sharp (eds). The 11th International Wheat Genetics Symposium proceedings. Sydney University Press, Sydney, Australia. ISBN: 978-1-920899-14-1. Available at http://hdl.handle.net/2123/3390, verified 6 Jan 2011.

Strelchenko P., K. Street, O. Mitrofanova, M. Mackay, K. Chabane, and J. Valkoun (2003). The genetic relationships between hexaploid wheat landraces from different geographical origin. p. 637-640. *In*: Proceeding from the 10th International Wheat Genetics Symposium (TIWGS), September 1-6, 2003, Paestum, Italy: vol 2. Available at http://www.cerealicoltura.it/simposyum/, verified 6 Jan 2011. [*cf.* Strelchenko et al., 2004]

Strelchenko, P., K. Street, O. Mitrofanova, M. Mackay, and F. Balfourier (2004). Genetic diversity among hexaploid wheat landraces with different geographical origins revealed by microsatellites: comparison with AFLP, and RAPD data. *In*: Fischer et al. New directions for a diverse planet, 4th International Crop Science Congress (4ICSC), 26 September to 1 October 2004, Brisbane, Australia. The Regional Institute Ltd, Gosford, Australia. ISBN: 1-920842-21-7. Available at http://cropscience.org.au/icsc2004/poster/3/3/1/940_strelchenkop.htm, and http://www.regional.org.au/au/asa/2004/poster/3/3/1/940_strelchenkop.htm, verified 6 Jan 2011.

Strelchenko P., K. Street, O. Mitrofanova, H. Hill, R. Henry, and M. Mackay (2008). Comparative assessment of wheat landraces from AWCC; ICARDA and VIR germplasm collections based on the analysis of SSR markers. *In*: Appels, R., R. Eastwood, E. Lagudah, P. Langridge, M. Mackay, L. McIntyre, and P. Sharp (eds). The 11th International Wheat Genetics Symposium proceedings. Sydney University Press, Sydney, Australia. ISBN: 978-1-920899-14-1. Available at http://hdl.handle.net/2123/3492, verified 6 Jan 2011.

Surowiecki, J. (2004). The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations. Abacus. ISBN 9780349116051. [Reprint 2005]

Tanksley, S.D. and S.R. McCouch (1997). Seed Banks and molecular maps: Unlocking genetic potential from the wild. Science 277(5329): 1063-1066. DOI: 10.1126/science.277.5329.1063.

Tannahill, R. (2002). Food in history. New and updated edition. Headline Book Publishing, London, UK. ISBN: 0-7472-6796-0. [First published in 1973]

TDWG (2005). Access to Biological Collections Data (ABCD), version 2.06 [Online]. ABCD task group, Biodiversity Information Standards (TDWG), hosted by the Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark; and the Natural History Museum (NHM), London, UK. Available at http://www.tdwg.org/standards/115/, verified 24 Jan 2011.

TDWG (2009). Darwin Core [Online]. Darwin Core task group, Biodiversity Information Standards (TDWG), hosted by the Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark; and the Natural History Museum (NHM), London, UK. Available at http://www.tdwg.org/standards/450/, verified 24 Jan 2011.

Tekauz, A. (1985). A numerical scale to classify reactions of barley to *Pyrenophora teres*. Canadian Journal of plant pathology 7: 181-183. Available at http://www.cps-scp.ca/download/cjpp-archive.1005/Vol7/CJPP7(2)181-183(1985).pdf, verified 25 Jan 2011.

Terrell, E.E. (1977). A checklist of names for 3,000 vascular plants of economic importance. USDA Agriculture Handbook 505. USDA, Washington DC, USA. 201 p. [Second edition in 1986, 241 p.]

Thachuk, C., José Crossa, J. Franco, S. Dreisigacker, M. Warburton, and G.F. Davenport (2009). Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measures. BMC Bioinformatics 10: 243. DOI: 10.1186/1471-2105-10-243. Core Hunter is available for download from: http://www.corehunter.org/, verified 25 Jan 2011.

Thanos, C.A. (1994). Aristotle and Theophrastus on plant-animal interactions. p. 3-11. *In*: Arianoutsou, M. and R.H. Groves (eds). Plant-animal interactions in the Mediterranean-type ecosystems. Kluwer Academic Publishers, Dordrecht.

Thompson, P.A. (1970). Seed banks as a means of improving the quality of seed lists. Taxon 9(1): 59-62.

Thormann, I., A. Lane, and D.T.F. Endresen (2007). Crop wild genetic resources, the CWR portal [Electronic Slides]. CWR IMC Meeting 28 August 2007, Bioversity International, Rome, Italy. Available at http://www.slideshare.net/DagEndresen/, verified 25 Jan 2011.

Tschermak, E. von (1900). *Über künstliche Kreuzung bei Pisum sativum*. Berichte der Deutschen botanischen Gesellschaft 18: 232-239. (Received for publication 2 June 1900).

Ullstrup, A.J. (1972). The impact of the southern corn leaf blight epidemics of 1970-71. Annual Review of Phytopathology 10: 37-50. DOI: 10.1146/annurev.py.10.090172.000345.

United Nations (1948). Protocol for the dissolution of the International Institute of Agriculture and the transfer of its functions and assets to the Food and Agriculture Organization of the United Nations [with Annex] Rome, 30th March 1946. Treaty series no 29. Available at http://www.fco.gov.uk/resources/en/pdf/treaties/TS1/1948/29, verified 4 Jan 2011.

United Nations (1993). Convention on Biological Diversity. Opened for signature at the Earth Summit 5 June 1992, Rio de Janeiro, Brazil. United Nations, New York, USA. Available at http://www.cbd.int/convention/text/ and at http://treaties.un.org/doc/Treaties/1992/06/19920605%2008-44%20PM/Ch_XXVII_08p.pdf, verified 29 Dec 2010.

United Nations (2002). Bonn guidelines on access to genetic resources and fair and equitable Sharing of the benefits arising out of their utilization. Secretariat of the Convention on Biological Diversity, Montreal, Canada. ISBN: 92-807-2255-7. Available at http://www.cbd.int/doc/publications/cbd-bonn-gdls-en.pdf, verified 30 Dec 2010.

United Nations (2004). World Population to 2300. United Nations, Department of Economic and Social Affairs, Population Division, New York, USA. 254 p. Available at http://www.un.org/esa/population/publications/longrange2/WorldPop2300final.pdf, verified 25 Jan 2011.

United Nations (2007). World Population Prospects, The 2006 Revision, Executive Summary. United Nations, Department of Economic and Social Affairs, Population Division, New York, USA. 19 p. URL: http://www.un.org/esa/population/publications/wpp2006/English.pdf, verified 25 Jan 2011.

United Nations (2010). Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity. Adopted at Nagoya on 29 October 2010. United Nations, New York, USA. Available at http://www.cbd.int/abs/ and at http://treaties.un.org/doc/Treaties/2010/11/20101127%2002-08%20PM/Ch-XXVII-8-b.pdf, verified 29 Dec 2010.

Upadhyaya, H.D., P.J. Bramel, and S. Singh (2001). Development of a chickpea core subset using geographic distribution and quantitative traits. Crop Science 41: 206-210. DOI: 10.2135/cropsci2001.411206x.

Upadhyaya, H.D. and R. Ortiz (2001). A mini core subset for capturing diversity and promoting utilization of chickpea genetic resources in crop improvement. Theoretical and Applied Genetics 102: 1292-1298. DOI: 10.1007/s00122-001-0556-y.

Upadhyaya, H.D., K.N. Reddy, M. Irshad Ahmed, and C.L.L. Gowda (2009). Identification of geographical gaps in the perl millet germplasm conserved at ICRISAT genebank from West and Central Africa. Plant Genetic Resources: Characterization and Utilization 8(1): 45-51. DOI: 10.1017/S147926210999013X.

UPOV (1991). International convention for the protection of new varieties of plants of December 2, 1961, as revised at Geneva on November 10, 1972, on October 23, 1978, and on March 19, 1991. International Union for the Protection of New Varieties of Plants (UPOV), Genève, Switzerland. Available at http://www.upov.int/en/publications/conventions/1991/act1991.htm, verified 25 Jan 2011.

USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network - (GRIN) [Online Database]. National Germplasm Resources Laboratory, Beltsville, Maryland. Available at http://www.ars-grin.gov/cgi-bin/npgs/html/index.pl, verified 25 January 2011.

Varmuza, K. and P. Filzmoser (2009). Introduction to multivariate statistical analysis in chemometrics. CRC Press, Boca Raton, NW, USA. ISBN: 978-1-4200-5947-2.

Vavilov, N.I. (1917). *O proiskhozhdenii kul'turnoy rzhi* [On the origin of cultivated rye]. *Tr. byuro po prikl. botan* [Bulletin of the Bureau of Applied Botany] 10(7-10): 561-590. [In Russian]

Vavilov, N.I. (1920). *Zakon gomologicheskikh ryadov v nasledstvennoy izmenchi vosti* [The law of homologous series in the case of hereditary variation]. Proceedings of the third All-Russian plant breeding conference. Saratov. 16 p. [In Russian]

Vavilov, N.I. (1922). The law of homologous series in variation. Journal of Genetics 12: 47-89. DOI: 10.1007/BF02983073.

Vavilov, N.I. (1923) [N.I. Vavilov correspondence archive, letter by dated 1923; *cf*. Esakov, 1980; *cf*. Cohen, 1982; *cf*. Loskutov, 1999:16]

Vavilov, N.I. (1924). *O vostochnykh tzentrakh proiskhozhdeniya kulturnykh rastenii* [On the Eastern centers of origin of cultivated plants]. *Noviy Vostok* [The New East] 6: 291-305. [In Russian, translation to English in Dorofeev, 1992: 1-13]

Vavilov, N.I. (1926). *Tsentry proiskhozhdeniya kul'turnykh rasteniy* [Centers of origin of cultivated plants]. *Tr. po prikl. botan i selek*. [Papers on applied botany and plant breeding] 16(2), 248 p. [In Russian, translation to English in Dorofeev, 1992: 22-135]

Vavilov, N.I. (1940). *Uchenie o proiskhozhdenii kul'turnykh rasteniy posle Darvina* [The theory of the origin of cultivated plants after Darwin]. *Sov. Nauka* [Soviet Science] 2: 55-75. [In Russian, translation to English in Dorofeev, 1992: 421-442]

Vavilov, N.I. (1957). *Mirovye resursy sortov khlebnykh zlakov, zernovykh bobovykh, l'na i ikh ispol'zovanie v selektsii. Opyt agroklimaticheskogo obozreniya vazhneyshikh polevykh kul'tur* [World resources of cereals, grain leguminous crops and flax and their utilization in plant breeding. General part. Agroecological survey of the principal field crops]. Izdatel'stvo Akademii Nauk SSR, Moskva, Leningrad. 463 p. [In Russian]

Vavilov, N.I. (1964). World resources of cereals, legumes, flax cultivars and their utilization in breeding. Wheat. (Nauka, Moscow and Leningrad). [In Russian] [*cf*. Strelchenko et al., 2004]

Vavilov, N.I. (1997). Five continents. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. ISBN: 92-9043-302-7. [Editor in chief: L.E. Rodin; Translated from Russian by Doris Löve; Edited by Reznik, S, and P. Stapleton]

Velleman, P.E. and L. Wilkinson (1993). Nominal, ordinal, interval, and ratio typologies are misleading. The American Statistician 47:65-72. Stable URL: http://www.jstor.org/stable/2684788, verified 2 Feb 2011.

Veteläinen M., V. Negri, and N. Maxted (eds) (2009). European landraces on-farm conservation, management and use. Bioversity Technical Bulletin No. 15. Bioversity International, Rome, Italy. ISBN: 978-92-9043-805-2.

Vicente, M.C. de and J.C. Glaszmann (eds) (2006). Molecular markers for allele mining. Proceedings of a workshop, 22-26 August 2005. MS Swaminathan Research Foundation, Chennai, India & International Plant Genetic Resources Institute (IPGRI), Rome, Italy. 85 p.

Vries, H. de (1889). Intracellular Pangenesis. Gustav Fischer Verlag, Jena, Germany. [English translation from German by C.S. Gager in 1910, Open Court Publishing Co., Chicago, USA. Available at http://www.esp.org/books/devries/pangenesis/facsimile/, verified 25 Jan 2011.

Vries, H. de (1900). *Das Spaltungsgesetz der Bastarde (Vorläfige Mittheilung).* Berichte der Deutschen botanischen Gesellschaft 18: 83-90. (Received for publication 14 March 1900). (*cf*. Roberts, 1929).

Ward, J.H., Jr. (1963), Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 48, 236-244. Stable URL: http://www.jstor.org/stable/2282967.

Weibull, J. (2010). *SKUD loggar ut 1 januari?* [SKUD offline from 1 Jan?] [Online]. Programmet för Odlad Mångfald (POM), Centrum för biologisk mångfald (CBM), Alnarp, Sweden. Available at http://www.pom.info/aktuellt/101216.htm, verified 25 Jan 2011. [In Swedish]

Weibull, P. (1997). *Växtförädlingsavgifter, växtförädlarrätt och patent*. p. 293-298. *In*: Olsson, G. (ed). Den svenska växtförädlingens historia. Jordbruksväxternas utveckling sedan 1880-talet. Skogs- och lantbrukshistoriska medelanden nr 20. Kungliga Skogs- och Lantbruksakademien, Stockholm, Sweden and SHS Text & Tryck, Hållsta, Sweden. 320 p. ISBN: 91-88780-32-5. [In Swedish]

Weismann, F.L.A. (1893). The Germ-Plasm, A theory of Heredity. Charles Scribners's Sons, New York, USA. Available at http://www.esp.org/books/weismann/germ-plasm/facsimile/, verified 2 Jan 2011.

Wiersema, J.H. and B. Léon (1999). World economic plants, a standard reference. CRC Press LLC, Boca Raton, Florida, USA. ISBN: 0-8493-2119-0.

Willner, E., N.R. Sackville Hamilton, and H. Knüpffer (1998). On the identification of duplicate accessions. p. 92-95. *In*: Maggioni, L., P. Marum, R. Sackville Hamilton, I. Thomas, T. Grass, and E. Lipman (compilers). Report of a working group on forages. Sixth meeting, 6-8 March 1997, Beitostølen, Norway. European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR), International Plant Genetic Resources Institute, Rome, Italy. ISBN: 92-9043-379-5.

Wilson, E.O. (1992). The diversity of life. Penguin Books, London, UK. 406 p. ISBN: 9780140169775.

Wise, B.M., N.B. Gallagher, R. Bro, J.M. Shaver, W. Windig, and R.S. Koch (2006). Chemometrics tutorial for PLS Toolbox and Solo. Eigenvector Research Inc., Wenatchee, WA, USA. ISBN: 0-9761184-1-6.

Wold, S. (1976). Pattern recognition by means of disjoint principal component models. Pattern Recognition 8: 127-139.

Wold, S. and M. Sjostrom (1977). SIMCA: A method for analyzing chemical data in terms of similarity and analogy. p. 243-282. *In*: Kowalski, B.R. (ed). Chemometrics Theory and Application, American Chemical Society Symposium, Series 52. American Chemical Society, Washington D.C., USA.

Wouw, M. van de, C. Kik, T. van Hintum, R. van Treuren, and B. Visser (2009). Genetic erosion in crops: concept, research results and challenges. Plant Genetic Resources: Characterization and Utilization 8(1): 1-15. DOI: 10.1017/S1479262109990062.

Wouw, M. van de, T. van Hintum, C. Kik, R. van Treuren, and B. Visser (2010). Genetic diversity trends in twentieth century crop cultivars: A Meta analysis. Theoretical and Applied Genetics 120(6): 1241-1252. DOI 10.1007/s00122-009-1252-6

WTO (1994). Agreement on trade-related aspects of intellectual property rights (TRIPS). Annex 1C of the Marrakesh agreement establishing the World Trade Organization (WTO), signed in Marrakesh, Morocco on 15 April 1994. Available at http://www.wto.org/english/docs_e/legal_e/27-trips_01_e.htm, verified 4 Dec 2010.

Wulff, E.V. (ed) (1935*). Kulturnaya Flora SSSR [Cultivated Plants of the USSR]. VIR, Moscow and Leningrad State Agricultural Publishing Company, Moscow-Leningrad, USSR (*cf*. Berg, 1950). [In

Russian, with Cyrillic characters]. * Volume 1: Grain cereals: Wheat (1935). 434 p. Edited by K.A. Flyaksberger and R.J. Rozhevits. Volume 2: Grain cereals (1936). 447 p. Part 1: Rye; Part 2: Barley; Part 3: Oat. Edited by V.F. Antropovy, A.I. Mordvinkina, and A.A. Orlov. Volume 3: Groat Crops: buckwheat, millet, rice. Volume 4: Grain legumes (1937). 680 p. Part 1: Pea; Part 2: Vetch. Edited by V.S. Muratova, F.L. Zalkind, L.I. Govorov, E.E. Ditmer, and N.R. Ivanov. Volume 5: Fiber crops (1940). 315 p. Edited by P.F. Medvedev and G.A. Pereverzev. Volume 6: Maize. Volume 7: Oleiferous Plants (1941). 496 p. Part 1: *Brassica, Camelina, Conringa*; Part 2: Perilla, Helianthus. Edited by A.I. Saltykovskij, N.I. Sarapov, A.A. Baburina, and F.S. Venclavovič. Volume 8: [No publication details for volume 8 found]. Volume 9: Potato (1941). Volume 10: Allium, Onion. Volume 11: Cabbage. Volume 12: Leafy Vegetables: asparagus, rhubarb, sorrel, spinach, purslane, garden cress, dill, chicory, and lettuce. Volume 13: Perennial Leguminous Herbs (1950). 526 p. Part 1: Alfalfa, Medic, Sweet clover, and Fenugreek; Part 2: Clover, Birds-foot trefoil, Trifolium, Lotus. Edited by E.N. Sinskaja, A. Lubenec, Platon, and M.I. Vavylov. Volume 14: Pome fruits: apple, pear, and quince. Volume 15: [No publication details for volume 15 found]. Volume 16: Small Fruit Crops: Berries cultures (1940). 285 p. Volume 17: Nut-bearing crops (1940). Volume 18: Root Crops: Fam. *Brassicaeae* - turnip, rutabaga, radish, and small radish. Volume 19: Root Crops: Fam. *Chenopodiaceae* - beet, Fam. *Umbelliferae* - carrot, parsley, celery, and parsnip (1950). Volume 20: Vegetable Plants: Fam. *Solanaceae* - tomato, common eggplant, black nightshade, pepino, pepper, husk tomato, and mandrake. (1950). 531 p. Edited by D.D. Brežnev. Volume 21: Part 1: Fam. *Cucurbitaceae* - watermelon, and pumpkin, Part 2: Fam. *Cucurbitaceae* - cucumber, and melon. Volume 22: [No publication details for volume 22 found].

Wulfsohn, D. (2010). Sampling techniques for plants and soil. p. 3-30. *In*: Upadhyaya, S., K. Giles, S. Haneklaus, and E. Schnug. Advanced engineering systems for speciality crops: A review of precision agriculture for water, chemical, and nutrient application and yield monitoring. Landbauforschung Völkenrode, Special Issue 340. vTI Agriculture and Forestry Research, Braunschweig, Germany. ISBN: 978-3-85675-066-1.

Yndgaard, F. (1982). A documentation system for the Nordic Gene Bank. FAO/IBPGR, Plant Genetic Resources Newsletter 49: 34-36.

Yndgaard, F. (1983). A procedure for packing long-term storage seed. Plant Genetic Resources Newsletter 54: 1-7.

Yndgaard, F. (1985). Genebank security storage in permafrost. Plant Genetic Resources Newsletter 62: 2-7.

Yndgaard, F. (1990). Computerized pedigree information on Nordic barley cultivars. Journal of the Swedish Seed Association 100(2): 105-111.

Yndgaard, F. and E. Kjellqvist (1982). *Orientering om Nordisk Genbank* [Introducing the Nordic Gene Bank]. Journal of the Swedish Seed Association, 92:75-86. [In Swedish with summary in English]

Yndgaard, F. and E. Kjellqvist (1984). Economic aspects of genebank conservation. Plant Genetic Resources Newsletter 58: 34-36.

Zeven, A.C. (1998). Landraces: A review of definitions and classifications. Euphytica 104(2): 127-139. DOI: 10.1023/A:1018683119237.

Zeyen, R.J. and J.V. Groth (2009). Sir Bent Skovmand: Champion for plant genetic resource preservation [Online]. ASPnet Features. DOI: 10.1094/ASPnetFeature-2010-0600. Available at http://www.apsnet.org/publications/apsnetfeatures/Pages/Skovmand.aspx, verified 19 Jan 2011.

Paper I

Endresen, Dag Terje Filip and Helmut Knüpffer (2011). The Darwin Core extension for genebanks opens up new opportunities for sharing genebank datasets. *Submitted to Biodiversity Informatics on 31 Jan 2011.*

The Darwin Core extension for genebanks opens up new opportunities for sharing genebank datasets

Dag Terje Filip Endresen * (1), Helmut Knüpffer (2)
(1) Nordic Genetic Resources Center, Alnarp, Sweden; (2) Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany. * Corresponding Author: dag.endresen@nordgen.org

ABSTRACT

Darwin Core (DwC) defines a standard set of core terms to describe the primary biodiversity data. Primary biodiversity data are data records derived from direct observation of species occurrences in nature or describing specimens in collections. The DwC can be seen as an extension to the standard Dublin Core metadata terms. The new DwC extension for genebanks (DwC-germplasm) defines additional terms required for describing genebank datasets, and is based on the established standards from the plant genetic resources community. The Global Biodiversity Information Facility (GBIF) provides an information infrastructure for biodiversity data including a suite of software tools for data publishing, distributed data access, and the capture of biodiversity data. The DwC-germplasm is a key component that provides access for the genebanks and the plant genetic resources community to the GBIF informatics infrastructure including the new toolkits for data exchange.

Abbreviations:
ABCD, Access to Biological Collections Data; *BioCASE*, Biological Collection Access Service for Europe; *CGIAR*, Consultative Group on International Agricultural Research; *COMECON*, The Council for Mutual Economic Assistance; *CWR*, Crop Wild Relatives; *CWRML*, Crop Wild Relative Markup Language; *DwC*, Darwin Core; *DwC-germplasm*, Darwin Core extension for genebanks; *EPGRIS*, European Plant Genetic Resources Information Infra-Structure; *EURISCO*, European Plant Genetic Resources Catalogue; *FAO*, Food and Agriculture Organization of the United Nations; *GBIF*, Global Biodiversity Information Facility; *GBIF IPT*, GBIF Integrated Publishing Toolkit; *GBRDS*, Global Biodiversity Resources Discovery System; *GCP*, Generation Challenge Programme; *IBPGR*, International Board for Plant Genetic Resources; *IPGRI*, International Plant Genetic Resources Institute; IPK, Leibniz Institute of Plant Genetics and Crop Plant Research (Gatersleben, Germany); *MCPD*, Multi-Crop Passport Descriptors; *NGB*, Nordic Gene Bank; *NordGen*, Nordic Genetic Resource Center; *PGR*, Plant Genetic Resources; *SESTO*, Seedstore documentation system; *SINGER*, System-wide Information Network for Genetic Resources; *TDWG*, Biodiversity Information Standards (former Taxonomic Databases Working Group).

A worldwide distributed network of genebanks

There are now more than 1750 genebanks distributed all around the world, with more than 130 large and medium-size genebank collections holding more than 10 000 accessions each (FAO, 2010). Each of these genebanks maintains living material of plant genetic resources. New accessions are added to the genebank collections from new collecting expeditions and from old cultivars obsolete to the commercial seed trade. The genebank documentation systems are continuously being extended with new accessions and with updated information on existing accessions. To build the global information system on Plant Genetic Resources for Food and Agriculture (PGRFA) described in the International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA; FAO, 2009:29, Article 17.1), the global information system needs to be frequently updated with new and updated data from each genebank collection (and other

information sources). With modern information technology, a distributed information system can be designed to allow extracting a snapshot of the distributed genebank datasets at any time. Moreover, a distributed germplasm information system will make updated germplasm information more accessible to plant breeders, crop scientists and other users, thus providing better and easier access to the plant material. The GBIF IPT provides one example of a software tool to build a distributed network of biodiversity databases. With the new Darwin Core extension for genebanks (DwC-germplasm), this new software tool is now available for the development of a distributed global information system for PGR collections.

ITPGRFA – International Treaty on Plant Genetic Resources for Food and Agriculture

The Global Plan of Action (FAO, 1996) was developed in 1996 at the International Technical Conference on Plant Genetic Resources in Leipzig, Germany, and set the scene for global collaboration on plant genetic resources in response to the Convention of Biological Diversity (CBD, 1993). The ITPGRFA was developed in 2001 (FAO, 2002, 2009) and entered into force on 29 June 2004. Article 17.1 of the ITPGRFA instructs the contracting parties to contribute to a distributed information system for PGRFA based on the exchange of information between existing information systems:

> *"The Contracting Parties shall cooperate to develop and strengthen a global information system to facilitate the exchange of information, based on existing information systems, on scientific, technical and environmental matters related to plant genetic resources for food and agriculture, with the expectation that such exchange of information will contribute to the sharing of benefits by making information on plant genetic resources for food and agriculture available to all Contracting Parties. In developing the Global Information System, cooperation will be sought with the Clearing House Mechanism of the Convention on Biological Diversity"* (FAO, 2002:22; FAO, 2009:29, Article 17.1).

The first genebank, the Vavilov Institute

The first genebanks for *ex situ* conservation of plant genetic resources were established more than one century ago, well before the advent of the digital computer. In 1894, professor A.F. Batalin, Director of the Sankt Petersburg Botanical Garden, made the initiative to organize the Bureau of Applied Botany under the Scientific Committee of the Russian Ministry of Agriculture. During 1901 and 1902, requests were distributed throughout the Russian provinces to collect and return seeds of local cultivars (landraces) of Russian agricultural crops. In 1908 the institute organized its first dedicated collecting expedition to collect Russian landraces (Regel, 1915; c.f. Loskutov, 1999). Nikolai Ivanovich Vavilov (1887-1943) joined the institute in 1910 and became its director in 1920. Under the leadership of Vavilov, the mandate of the institute was expanded to include the long-term conservation of plant genetic resources. This marks the advent of the modern genebanks, as we know them today (Loskutov, 1999). The Bureau of Applied Botany was the predecessor to the current NI Vavilov Institute for Plant Industry (VIR). The first online genebank database for VIR was hosted in Bonn, Germany, during 1994 to 1999, in collaboration with the Information Centre for Genetic Resources (IGR/ZADI; now Information Centre for Biological Diversity, IBV/BLE) (Harrer and Omelchenko, 1998). In 2000, the online database for VIR was moved to a new web server within the institute in Sankt Petersburg and published at http://www.vir.nw.ru (verified 31 Jan 2011) (Omelchenko et al., 2003).

The first genebank information systems, *Index Seminum*

The seedbanks of botanical gardens and their seed exchange system can be seen as a predecessor to the present genebanks. Around 1543, the first botanical gardens in Europe were established in Italy (Stafleu, 1969; Stearn, 1971). The botanical gardens have traditionally published seed lists (*Index Seminum*) for the purpose of seed exchange (Heywood, 1964). However the seed exchange of the botanical gardens has been criticized for problems with

inaccurate classification, poor viability and the lack of information on the origin of the seeds (Thompson, 1970). The aforementioned Bureau of Applied Botany in Russia included, from its start in 1894, an information department with the task to provide information on the availability of seed from both cultivated and wild species (Loskutov, 1999). The last seed catalog of VIR (*Delectus Seminum* [list of selected seeds]) was published in 1999 (Dragavtsev et al., 1999). The crop departments continued, however, to produce more detailed crop-based "catalogues" after the last *Delectus Seminum* (see for example Loskutov and Ryabchenko, 2002). In Germany, Rudolf Mansfeld published the first *Index Seminum* issued by the Gatersleben "World Collection of Cultivated Plants" in 1947 (Hammer, 2003). The last *Index Seminum Gaterslebensis* was published in two parts in 1999 for 2000 (Knüpffer, 1999a, b). An online catalogue has been operational since 1996 and made the Index Seminum with its total of 413 pages, distributed to about one thousand recipients, superfluous. The Nordic Gene Bank was from its beginning instructed to produce a printed *Index Seminum* (Palmstierna et al., 1975:48). The last printed seed catalog was published in 1989 (NGB, 1989). The genebank search database at Nordic Gene Bank (NGB) has been available online since 1993 (Huldén et al., 1998).

The first electronic genebank information systems

The Fifth Yugoslav Symposium on Research in Wheat in 1966 included one of the first initiatives to develop international standards and mechanisms for sharing electronic documentation of crop genetic resources (Konzak and Sigurbjörnsson, 1966). The Food and Agriculture Organization of the United Nations (FAO) and the International Atomic Energy Agency (IAEA) assembled a group of experts in Vienna. This expert group proposed to establish a distributed network with national crop information centers reporting national crop data to a central hub to be set up at FAO in Rome, Italy. The central file maintained and hosted at FAO would be published to become available for plant breeders, crop research and policy organs (Finlay and Konzak, 1970). When the IBPGR (International Board for Plant Genetic Resources) was established in 1974, one of its first tasks was to organize and coordinate a distributed network of genebank information systems. Work was initiated on a distributed system under the name Genetic Resources Communication, Information and Documentation System (GR/CIDS) (IBPGR, 1976, 1977).

> *"The central file at FAO will fulfill two main roles. First, by accepting records from all holders of collections willing to exchange seed and, by adding information on new material, it will become a current record of stocks available throughout the world. Second, by also accumulating information about material which has severely restricted seed supply or for which seed is not available, the central file will become an archive for records of genetic variation"* (Finlay and Konzak, 1970:463-464).

> *"When it is fully developed, GR/CIDS should encompass the whole of the information component, including documentation and the flow of information, of genetic resources work, from the initial collection of data about traditional materials in the field to the performance of improved varieties derived from them"* (IBPGR, 1976:5).

A global accession-based genebank information system is still not completed. Its development, and the sharing of characterization and evaluation (C&E) data remain today a high priority for a rational utilization of plant genetic resources for food and agriculture (FAO, 2010).

> *"The first SoW [State of the World] report highlighted the poor documentation availability on most of the world's ex situ PGR. This problem continues to be a substantial obstacle to the increased use of PGRFA in crop improvement and research. Where documentation and characterization data do exist, there are frequent problems in standardization and accessibility, even for basic passport information"* (FAO, 2010:77).

Crop descriptor lists

Bioversity International (IBPGR 1974-1991; IPGRI 1991-2006) has developed and published more than 100 crop-specific descriptor lists since 1977 (Gotor et al., 2008). The first descriptor list was for cultivated potato (IBPGR, 1977) followed by the descriptors for wheat and *Aegilops* (IBPGR, 1978). These crop descriptor lists provided a valuable standard for documentation of PGR and in particular for their characterization and evaluation data (C&E) (Gator et al, 2008). The multi-crop passport descriptors (MCPD) were introduced in 1996 (Hazekamp et al., 1997) and published in the current format in 2001 (Alercia et al., 2001). In the context of the joint EURISCO search portal for genebanks in Europe, a few amendments were made to the MCPD including descriptive name of institutes, URL for linking to additional data. The EURISCO amendments included also the status of the accession in the multilateral system (MLS) of the ITPGRFA and in respect to the European Genebank Integrated System (AEGIS) (EURISCO, 2011). The genebanks have tried to follow these standard crop descriptors in projects to describe the genebank collections. This ensured a good interoperability between the genebank datasets from different institutes and countries. A new revision of the MCPD standard was recently announced at the Crop Genebank Knowledge Base Blog (SGRP, 2011).

In the former Eastern Bloc countries, the COMECON (Council for Mutual Economic Assistance, 1949-1991) crop descriptors ensured a similar standardization and interoperability of germplasm datasets. The first COMECON crop descriptor lists were released in 1974 for *Triticum* (Bareš, 1974), *Hordeum* and *Avena* (for an overview, see Knüpffer 1983). Their predecessors have been national crop descriptor lists of the USSR and Czechoslovakia since the 1960s (cf. Knüpffer 1983). A first standard for passport data recording across genebanks has been proposed by the COMECON PGR documentation working-group (Rogalewicz, 1988). The COMECON passport descriptor list is similar in aim and coverage as the later MCPD.
The crop descriptor lists from Bioversity International and the COMECON have contributed to acceptable data interoperability between the distributed genebank datasets across the world and make the present development of an automatic data exchange mechanism for PGR data easier.

Distributed genebank documentation systems in the Nordic countries

The Nordic Gene Bank (NGB) was established in January 1979 (Yndgaard and Kjellqvist, 1982; Ellerström, 1982). NGB was implemented as a regional center for Denmark, Finland, Iceland, Norway, and Sweden, to include the Nordic seed collections already established at that time, and to take over the function of national genebanks that would otherwise be needed (Palmstierna et al., 1975). An important task for the new institute was to interact with the national networks for conservation and utilization of plant genetic resources across the Nordic countries and to bring them together. It was soon realized that an efficient distributed information infrastructure was required to serve the Nordic crop networks and working groups (Yndgaard, 1981; Yndgaard and Kjellqvist, 1982). An external expert group was formed to develop the technical requirements for this distributed information network to work (Lindeberg et al., 1981). NGB was thus designed to form the central hub in this distributed information infrastructure with a national so-called Information Retrieval System (IRS) established in each of the Nordic countries. The NGB IRS-83 has been a CP/M based microcomputer with file management software. The NGB purchased 16 PCs to be configured as IRS-83 and distributed in 1983 nine IRS-83 to partner institutions across the Nordic countries for data capture and upload of data to NGB (Bjarnason, 1989). The IRS-83 would share data with NGB using a modem connected to the public phone line or by sending floppy discs by the public postal system (Lindeberg, 1981). The first database management system for the Nordic Gene Bank was developed in 1983 and named Biological Information Retrieval System (BIRS; Rydström, 1989). In 1989 the BIRS system was migrated as a module to the Nordic Biometry Information System (NOBIS) platform of the Nordic Biometry Project (NBP). Crop breeders and researchers across the

Nordic region used the BIRS and the NOBIS system. BIRS and NOBIS thus provide an early example of a distributed PGR data network in the Nordic countries (NGB, 1991). During the 1990s the NGB documentation migrated via the dBASE IV platform to Visual dBASE (Huldén et al., 1997, 1998). In January 2003, the NGB documentation system (SESTO) was upgraded to become a web application based on the PostgreSQL database backend and a front-end developed with the PHP scripting language (Endresen et al., 2005b). The transformation of the SESTO system into an online database allowed users across the Nordic countries to log in remotely and maintain their own datasets. In 2005, colleagues from the Baltic genebanks in Estonia, Latvia and Lithuania started to use SESTO to update their own genebank datasets (Endresen, 2005a; Jõgeva PBI, 2005). More recent on-site installations of SESTO for genebanks, respectively in Southeastern Europe and Eastern Africa (http://www.eapgren.net, verified 31 Jan 2011), have assisted the development of regional PGR networks in these regions (Endresen et al., 2005a; Endresen et al., 2005b; Thörn, 2006). The documentation of PGR in the Nordic countries is being carried out in a multi-national distributed network. One of the main roles of the information system at the Nordic Gene Bank (now NordGen) has been, since the very first origin, to integrate germplasm (passport) information of the genebank accessions in the central seed storage in Alnarp with performance data (characterization and evaluation data) received from plant breeders and agricultural research stations across the Nordic countries. The distributed Nordic PGR network will benefit substantially from further developments in standards and automatic exchange mechanisms for PGR data.

European Central Crop Databases (ECPGR)

During the 1980s and 1990s, a number of European Central Crop Databases (ECCDB) was developed as part of the European Cooperative Programme for Plant Genetic Resources (ECPGR) (Knüpffer, 1995; Lipman et al., 1997). The first such database was that of rye, developed by the Polish genebank at the Plant Breeding and Acclimatization Institute (IHAR). As the first of its kind (initiated in September 1981 at a joint meeting between the Polish gene bank and the Nordic Gene Bank), the rye catalogue comprised passport data of rye accessions maintained in 11 genetic resources centers (Serwiński and Konopka, 1984). This pioneer work was used as a reference as well as a model for other European databases (Podyma, 2001). The "European Barley List" (Knüpffer, 1987) published from the European Barley Database (Knüpffer, 1988), provided passport data of more than 55,000 accessions. The development of the European *Prunus* Database (EPDB) was initiated at NGB in 1983. The European *Prunus* Database was compiled in 1984 and made available electronically in dBASE format (Kurki, 1986). In 1989, the EPDB was published as a series of 5 printed catalogues (cherry, apricot, almond, peach, and plum) (Bjarnason and Niklasson, 1989; Niklasson and Bjarnason, 1989a, 1989b; Niklasson, 1989a, 1989b, 1989c). NGB published the European *Phleum* Database online in October 1996 (Huldén, 1997), and soon thereafter the European *Agrostis* and *Phalaris* Database followed. The ECCDBs contributed to European collaboration and joint project activities for PGR. The ECCDBs have also contributed to the mobilization of some characterization and evaluation data (C&E), but to a much smaller extent than presumed (Maggioni, 2007). Only 16 out of the 62 ECCDBs have C&E data online (Maggioni, 2009; http://www.ecpgr.cgiar.org/Databases/Databases.htm, verified 31 Jan 2011). The Central Crop Databases in Europe provide a distributed network of crop experts and publish an aggregated database with regular updates of data from the genebanks holding accessions of the respective crops. These crop networks would greatly benefit from a more standardized automatic data exchange mechanism.

EURISCO – European Search Catalogue for Plant Genetic Resources

In September 2003, EURISCO was released as a searchable online database with passport data from many European genebank collections (EURISCO, 2003; IPGRI, 2003). EURISCO was

developed during 2000-2003 by the EU-funded EPGRIS (Establishment of a European Plant Genetic Resources Information Infra-Structure) project (IPGRI, 2001, 2002). EURISCO is hosted by Bioversity International in Rome and is regularly updated by the designated national focal points representing almost all European countries (EURISCO, 2002). The EURISCO framework has been proposed as a model for other regions. One example is the presentation of the EURISCO infrastructure as a proposed model for the development of a distributed genebank network in Latin America (Gaiji et al., 2008). The many national and regional genebank institutions in Europe maintain numerous distributed genebank databases. Direct data exchange between each of these genebanks and the Central Crop Databases as well as with the EURISCO Catalogue (via the respective National Inventories) makes for a complex information infrastructure. There is already today a substantial flow of data through this germplasm information web and many man-hours dedicated to keep the data pathways open. The development of standardized and automatic data exchange mechanisms for PGR in Europe has been on the agenda of the ECPGR Documentation and Information Network and its predecessors, e.g. the Internet Advisory Group, for many years (ECP/GR, 1997, Maggioni, 2005, 2010).

PGR Forum, CWRIS, CWRML

The EU-funded PGR Forum project (2003-2005) (http://pgrforum.org, verified 31 Jan 2011) produced a new information system for Crop Wild Relatives (CWRIS) and a new XML (extensible markup language) based schema for exchange of datasets on crop wild relatives. The Crop Wild Relative Markup Language (CWRML) was designed for compatibility with the Darwin Core and the Access to Biological Collections Data (ABCD) data standards from Biodiversity Information Standards (TDWG). It was further envisioned that terms from the CWRML schema could be extracted to form an extension to the Darwin Core standard (Moore et al., 2008).

Genetic Resource Information Network (GRIN)

Efficient germplasm data management is the key to continued success in plant breeding to ensure future improved crop performance. The genebanks are faced with increasing new challenges of documenting new types of germplasm information including in particular molecular genetic descriptions (Tanksley and McCouch, 1997; Richards and Volk, 2010). In 2008 a collaborative project between the National Plant Germplasm System (NPGS) of the USDA-ARS, the Global Crop Diversity Trust and Bioversity International was initiated to develop a new genebank management software tool (Arnaud et al., 2008; Postman et al., 2010). The new software is named GRIN-Global and can be seen as a new, scalable version of the GRIN. GRIN was developed and has been further enhanced by the NPGS since 1984 (Anonymous, 1983, 1984). The development of pcGRIN was initiated in 1995 (and officially released in 1998) with a similar purpose as the new GRIN-Global; to provide a complete data management package for small genebanks and breeders based in the GRIN information system (Stearn, 1998). There was also an earlier version named 'PC-GRIN' released in 1993 that allowed for offline queries in crop-based subsets from the GRIN database (Greene, 1993). The GRIN-Global can thus be seen as an "upgrade" of the pcGRIN (but is of course not based on the source code from pcGRIN). The GRIN-Global information system will support data exchange using the TDWG standards and the GBIF infrastructure (FAO ITPGRFA, 2008:15).

International Crop Information System (ICIS)

The International Crop Information System (ICIS) was designed to integrate different types of germplasm information, mainly from the Consultative Group on International Agricultural Research (CGIAR) centers and their collaborators. The initial developments of ICIS started at the International Maize and Wheat Improvement Center (CIMMYT) in Mexico during the

1980s to integrate germplasm information on wheat (IWIS, International Wheat Information System, http://iwis.cimmyt.org/, verified 31 Jan 2011; Fox et al., 1996) from different sources with the genebank accessions in the CIMMYT genebank collection (Fox and Skovmand, 1996). Recent developments of ICIS for rice coordinated by the International Rice Research Institute (IRRI; http://www.iris.irri.org/, verified 31 Jan 2011) enhanced the capabilities of ICIS to integrate different germplasm datasets including the use of modern data exchange mechanisms to share germplasm datasets within the PGR community (Bruskiewich et al., 2003; McLaren et al., 2005; DeLacy et al., 2009).

Generation Challenge Programme (GCP)

The GCP is a time-bound (2003-2013) initiative of the CGIAR to use genetic diversity for improving the crops for resource-poor farmers in developing countries (http://www.generationcp.org, verified 21 Jan 2011). The GCP includes a theme for advanced research on crop information systems and bioinformatics (Bruskiewich et al., 2006; Bruskiewich et al., 2008). During 2005, the GCP Passport XML schema was harmonized for interoperability with the ABCD schema and processed for implementation with the Biological Collection Access Service for Europe (BioCASE) toolkit. The GCP Central Registry (http://gcpcr.grinfo.net, verified 31 Jan 2011) was designed to interact with research datasets published by the project partners using the BioCASE data publishing toolkit.

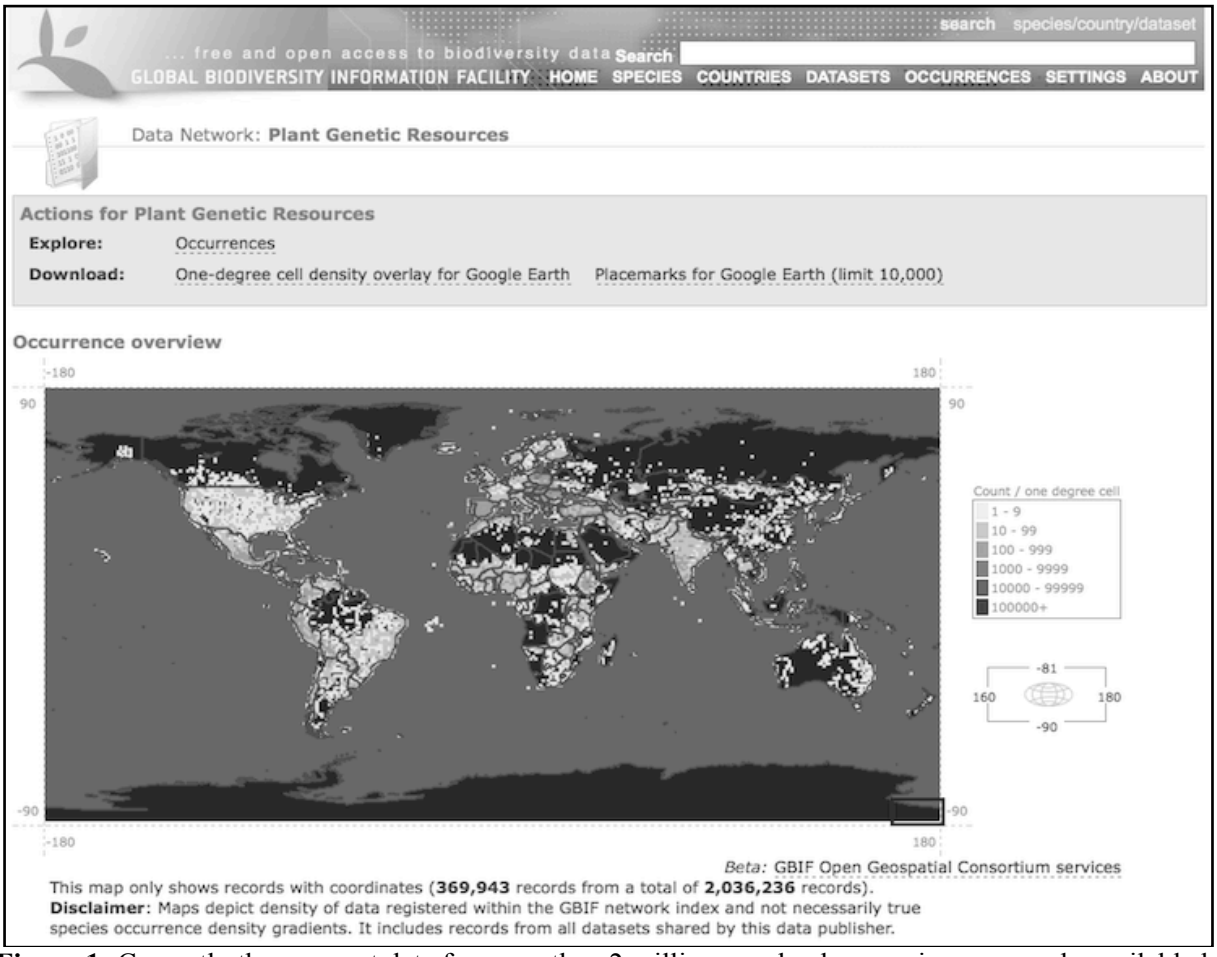


Figure 1: Currently the passport data for more than 2 million genebank accessions are made available by genebanks through the GBIF distributed data infrastructure. GBIF Data Portal showing the Plant Genetic Resources Data Network (http://data.gbif.org/datasets/network/2/, visited 31 Jan 2011).

The first genebanks to join the Global Biodiversity Information Network

The Global Biodiversity Information Facility (GBIF) was established in 2001 (GBIF, 2001) as an inter-governmental initiative to facilitate free and open access to biodiversity data online (http://www.gbif.org, verified 31 Jan 2011). During 2004, the Nordic Gene Bank, the Polish genebank in Radzików, and the German genebank in Gatersleben were the first genebanks to join the GBIF network (Knüpffer et al., 2004). The GBIF is a distributed biodiversity information network based largely on the information standards defined by the Biodiversity Information Standards organization (TDWG; Taxonomic Databases Working Group, 1985-2006). TDWG published in 2004 two alternative biodiversity collection data standards potentially suitable for genebank accessions, the Darwin Core (version 2) and the Access to Biological Collections Data (ABCD version 1.20) (Berendsohn, 2005). In collaboration with the ABCD task group, during 2005 the standard genebank descriptors (MCPD) were mapped to the corresponding ABCD terms, or added as new descriptors to an updated version of the ABCD (version 2.06) (Berendsohn and Knüpffer, 2006). This data interoperability between the crop datasets and the TDWG data-sharing standards opened the possibility for utilization of the GBIF data infrastructure by the PGR community for its own interoperability tasks. It was now possible to start the implementation of the data-sharing toolkits from the GBIF and the TDWG communities in the EURISCO network for Europe (Endresen et al., 2006). The Biological Collection Access Service for Europe (BioCASE; Berendsohn, 2002) data publishing toolkit was installed at 15 genebanks worldwide and a demo data portal was developed in 2006 to interact with these distributed web services established by the BioCASE installations (http://chm.grinfo.net/, verified 31 Jan 2011).

Dublin Core Metadata Initiative (DCMI)

The DCMI was initiated at a joint workshop between the Online Computer Library Center (OCLC) and the National Center for Supercomputing Applications (NCSA) on metadata semantics held in Dublin (Ohio, USA) in March 1995. The output from this workshop was called "Dublin Core metadata" based on the location of the workshop. The original target was to develop a small, common set of metadata elements to describe Web content. The original Dublin Core elements (or terms) were: Subject, Title, Author, Publisher, OtherAgent, Date, ObjectType, Form, Identifier, Relation, Source, Language, and Coverage. The Dublin Core was designed to be extensible (Weibel et al., 1995).

Darwin Core

The Natural History Museums in the USA started early to develop information networks with distributed query systems using the Internet. The Species Analyst project was initiated in 1997 and coordinated from Kansas University (Vieglais et al., 1998; Townsend Peterson et al., 2003). The first version of the Darwin Core standard was developed in 1999 by the Species Analyst project (Stein and Wieczorek, 2004). The Mammal Networked Information System (MaNIS) was initiated in 1999 and established in June 2002 a distributed information network between 17 North American mammal natural history collections. The MaNIS network was developed in parallel with the Distributed Generic Information Retrieval (DiGIR) data publishing toolkit and contributed to the development of the next version of the Darwin Core standard (DwC version 1.21) (Stein and Wieczorek, 2004). The current version of the Darwin Core is more different from the previous versions than the previous versions are from each other, and was ratified by TDWG in October 2010 (http://www.tdwg.org/standards/450/, verified 31 Jan 2011).

RESULTS

During 2008, work was started at GBIF for a major upgrade of the data publishing toolkit for the GBIF network. The new tool was named GBIF Integrated Publishing Toolkit (IPT) and is based

on the Darwin Core data model. The Darwin Core (DwC) standard was at the time under revision by TDWG for a new version to be ready during 2009. While the ABCD standard is very comprehensive with several thousand terms, the Darwin Core standard implements a more limited set of core terms with domain specific terms organized in a number of published extensions. There was, as of 2008, no Darwin Core extension to ensure full interoperability with the genebank information requirements. During a Darwin Core workshop in Copenhagen (hosted by GBIF) in January 2009, work was initiated to develop an extension for germplasm to the new revised Darwin Core standard. The Darwin Core extension for genebanks (DwC-germplasm) (http://rs.nordgen.org/dwc/, verified 31 Jan 2011) is required for the rational use of the GBIF IPT in the genebank community (Endresen et al., 2009). The DwC-germplasm provides a similar piece in the interoperability puzzle as the ABCD version 2.06 provided in 2005 to make rational use of the BioCASE toolkit in the genebank community possible.

Darwin Core extension for genebanks (DwC-germplasm)

The DwC-germplasm version 0.1 includes in the first draft version the following terms: GermplasmID, BiologicalStatusOfSample, BiologicalStatusOfSampleCode, GermplasmIdentifier, CollectingInstituteCode, AncestralData, PurdyPedigree, TypeOfStorage, LocationOfSafetyDuplication, SafetyDuplicationID, SafetyDuplicationInstituteCode, SafetyDuplicationInstitute, SafetyDuplicationDate, GermplasmTreatiesAndRegulations, GermplasmRegulationID, TreatyOrRegulationName, TreatyOrRegulationGoverningBody, SampleAcquisition, SampleAcquisitionID, DonorsSampleIdentifier, SampleAcquisitionSource, SampleAcquisitionDate, DonorInstituteCode, DonorInstitute, SampleAcquisitionRemarks, BreedingEventGroup, BreedingEventID, BreedersSampleIdentifier, BreedingYear, BreederPerson, BreederInstituteCode, BreederInstitute, BreedingCountry, BreedingCountryCode, BreedingEventRemarks, GermplasmTraitGroup, GermplasmTraitID, MeasurementByInstituteCode, MeasurementGrowthStage, GermplasmTraitIdentifier, GermplasmTraitClass, GermplasmTraitScale, GermplasmTraitSource, GermplasmTraitRemarks, GermplasmExperimentGroup, GermplasmExperimentID, GermplasmExperimentIdentifier, GermplasmExperimentRemarks, GermplasmExperimentYear, GermplasmExperimentReport. Definitions of these terms can be found in the namespace: http://rs.nordgen.org/dwc/germplasm/0.1/terms/.

The first draft version of the DwC-germplasm was published for discussion at the EPGRIS3 wiki (http://www.nordgen.org/epgris3/wiki/index.php/DwC_Germplasm, verified 31 Jan 2011). The EPGRIS3 (Establishment of a European Plant Genetic Resources Information Infra-Structure, phase 3) is an initiative of the ECPGR Documentation and Information Network. The initial development of the DwC-germplasm at the EPGRIS3 wiki attracted feedback and suggestions from the ENSCONET (European Native Seed Conservation Network) project regarding additional terms for *in situ* conservation. The Millennium Seed Bank proposed additional terms to describe *ex situ* conservation management. After receiving even further feedback from other communities outside the European genebank community, a new DwC-germplasm project home page was established at Google Code (http://code.google.com/p/darwincore-germplasm/, verified 31 Jan 2011). Further modifications and additional terms to the DwC-germplasm will be discussed and agreed here before they will be passed on to be consolidated within the genebank community and eventually included in the official version of the extension.

The official version of the DwC-germplasm extension is available from the GBIF vocabularies site (http://vocabularies.gbif.org/node/126981, verified 31 Jan 2011). Consolidated modifications to the DwC-germplasm will be implemented at this site. The GBIF vocabularies system will publish updates and modifications to the Darwin Core extensions to the GBRDS

(Global Biodiversity Resources Discovery System). The GBIF IPT is designed to communicate directly with the GBRDS to download new and updated Darwin Core extensions and vocabularies.

Deployment of the DwC-germplasm in the GBIF IPT

GBIF, NordGen and Bioversity International initiated, in 2010, a feasibility study to evaluate how the GBIF infrastructure can meet the needs of the European genebank community (Gaiji et al., 2010). This feasibility study was also coordinated with the ECPGR Documentation and Information Network for the genebank community in Europe (Maggioni, 2010). The prototype GBIF Integrated Publishing Toolkit (IPT version 1.0) was during 2010 installed in five genebanks within the European Plant Genetic Resources Catalogue (EURISCO) using the DwC-germplasm extension. These were the national genebanks in the Russian Federation (Vavilov Institute, Sankt Petersburg), Germany (IPK Gatersleben), Czech Republic (Crop Research Institute, Prague), the Netherlands (Wageningen University and Research Center, Centre for Genetic Resources), and within the Nordic and Baltic countries (genebank database hosted from the Nordic Genetic Resources Center). During the first months of 2011, two additional installations are scheduled at Bioversity International in Montpellier, France, and at the EURISCO headquarter at Bioversity International in Rome, Italy (Figure 2).

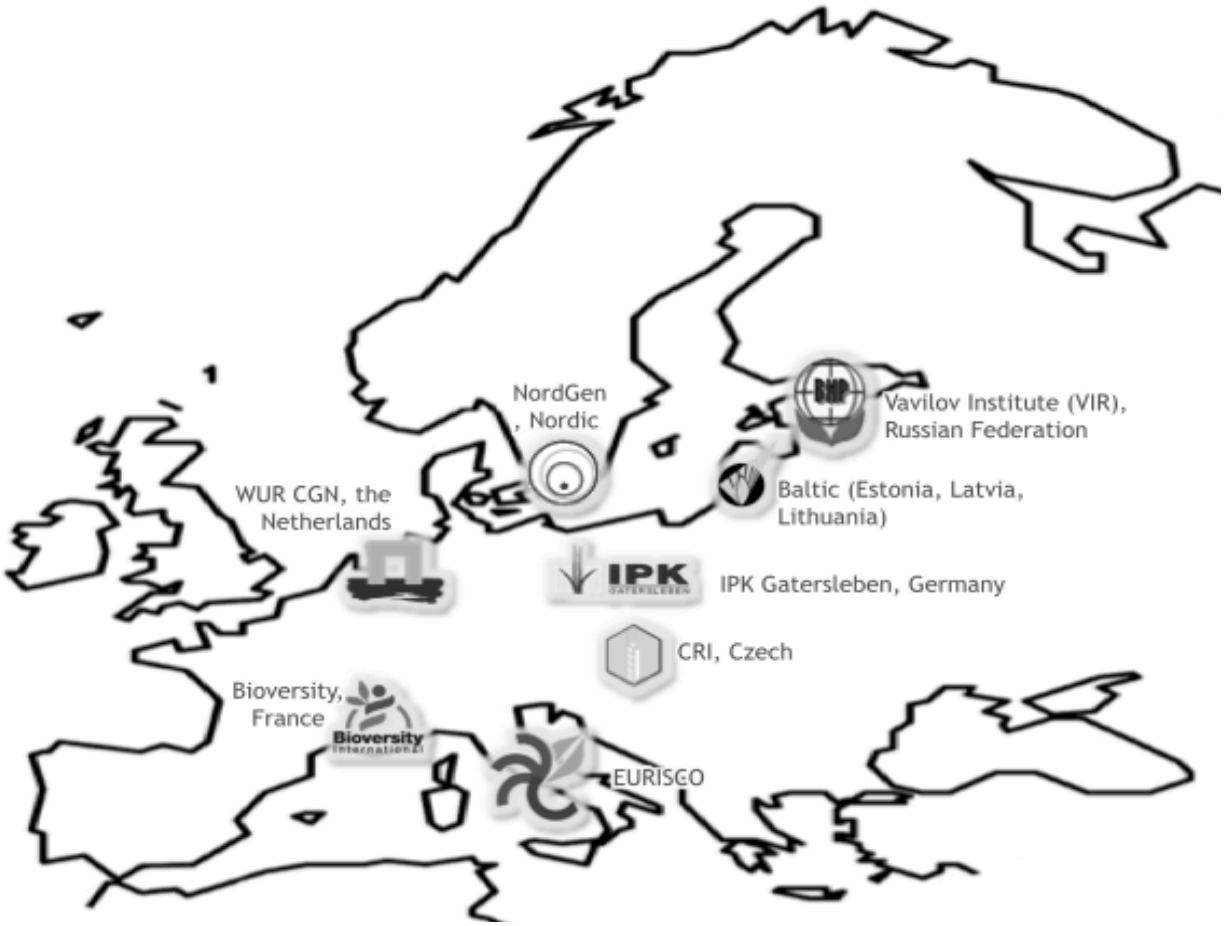


Figure 2: During the 2010 feasibility study for the European genebank community, the prototype GBIF Integrated Publishing Toolkit (IPT) was installed at the national genebanks in the Russian Federation, Germany, Czech Republic, the Netherlands, and at the Nordic Genetic Resource Center (hosting the genebank database for the Nordic and Baltic countries). During the first months of 2011, the GBIF IPT will also be installed at Bioversity International in Montpellier (France) and in Rome (Italy) to complete this feasibility study.

While the prototype version of the software caused some problems of instability, the installation (1), followed by the mapping of the genebank datasets to Darwin Core, including the gene bank extension (2) and finally the registration to the GBRDS (3) was completed satisfactorily. The hardware requirements and in particular the demands for internal memory were a major barrier encountered during most of the installations. The hardware requirements for the next version of the IPT software (version 2.0) have been significantly reduced. The experiences from the genebank feasibility study provided the IPT development team at GBIF with feedback and suggested improvements leading from the different prototype versions to the new version 2, soon to be released. It is important that the final release of the IPT software does not cause any instabilities to the Java application server it is installed into. Generally, the project resulted in a positive evaluation and the genebank community has started the initial plans for a second feasibility study to evaluate the IPT version 2.0 at other genebanks in Europe.

DISCUSSION

The Darwin Core standard is itself an extension of another standard, the Dublin Core Metadata Initiative (DCMI; http://dublincore.org/, verified 31 Jan 2011). The Dublin Core provides a bridge to ensure low-level interoperability between wide ranges of metadata standards. Implementing Darwin Core as an extension to the Dublin Core promotes the interoperability of biodiversity data with information from other domains. The Darwin Core set of core terms includes some terms from the Dublin Core 'terminology'. But more important than the shared terms, is the shared framework to describe and implement the terms in applied solutions.
The same principles apply to the benefits of building a genebank extension based on the Darwin Core, or adapting other solutions from outside the PGR network. By following a few ways and guidelines for 'best practices', the genebanks can with few efforts adapt tools and principles developed in other communities for efficient use in their own information network (Knüpffer et al., 2007).

> *"The achieved compatibility of data standards between PGR and biodiversity collections allows integrating the worldwide germplasm collections into biodiversity information networks. Using GBIF technology (and contributing to its development), the PGR community can easily establish specific PGR information networks without creating its own technology"* (Knüpffer et al., in press).

By finding a few common ways and guidelines for 'best practices', the genebanks can with fewer efforts adapt tools and principles developed in other parts of their 'own' community for efficient use across the entire PGR information network.

Automatic data exchange mechanisms

Many of the present data exchange mechanisms in use in the genebank community rely on laborious and repeated transformations of the original genebank datasets into the agreed standard formats. The genebanks in Europe regularly produce an updated subset from their information system complying with the EURISCO data exchange format (based on the multi-crop passport descriptors (MCPD)). Then the subset from each genebank in a country is combined into a so-called National Inventory and uploaded to the central EURISCO data portal (hosted by Bioversity International). The European Central Crop Databases (ECCDB) also request from each genebank to extract a similar subset from their information system. Many ECCDBs ask for data on selected descriptors from the Bioversity Crop Descriptor lists in addition to the MCPD. The ECCDBs are limited in scope each to a different crop species and ask thus for a different set of additional crop-specific descriptors. While the EURISCO has implemented an online data upload tool to receive the updated national inventories, the updated subsets for the ECCDBs are often exchanged as email attachments. The CGIAR genebanks share similar subsets from their information systems with the System-wide Information Network

for Genetic Resources (SINGER; http://singer.cgiar.org/, verified 31 Jan 2011). The FAO WIEWS (World Information and Early Warning System on PGRFA; http://apps3.fao.org/wiews/wiews.jsp, verified 31 Jan 2011) also requests, on a regular basis, updated subsets from all genebank worldwide. The requested format for these subsets is also roughly based on the MCPD standard. The record level data unit is however different, as WIEWS request metadata on stratified groups of genebank accessions, rather than the accession level data requested by EURISCO, ECCDBs and SINGER.

New data exchange mechanisms using web services have the potential to make all these aforementioned data exchange operations fully automatic. And with the new data provider toolkit software packages provided as an open source public good from the Global Biodiversity Information Facility (GBIF), the required efforts to establish and maintain such fully automatic multi-purpose data flow pathways with web services are getting less demanding and becoming more low-tech to implement. The GBIF Integrated Publishing Toolkit (IPT) is the latest and most user-friendly software package for sharing biodiversity datasets (such as the genebank datasets). The Darwin Core extension for genebanks (DwC-germplasm) provides a necessary 'plug-in' to make the new GBIF IPT available for rational use in the genebank community.
The genebank community was one of the first biodiversity information networks to develop this type of plug-in to start using the GBIF IPT. It is expected that the experiences from the development and implementation of the DwC-germplasm for the genebanks can provide some examples for other biodiversity information networks to study. With the following section, we aim to describe the steps to follow in order to develop a similar Darwin Core extension in other biodiversity information networks.

HOW TO create a new Darwin Core extension

The development and implementation of the Darwin Core extension for germplasm can be used as an example for other biodiversity information communities to develop their own DwC extensions. The following steps have to be carried out:

(1) The community needs to compile a consolidated list of terms to describe their data domain.

(2) After finding agreement on the terms with the relevant stakeholders inside the relevant community, these terms should be harmonized and mapped to the standard Darwin Core terms (http://rs.tdwg.org/dwc/terms/index.htm, verified 31 Jan 2011). New terms should only be defined for an extension if they are not already included in the standard core terms. Some of the descriptor terms implemented in a community may be similar to one of the core terms, but with a different formatting or a slightly different semantic meaning. Whenever possible, it is recommended to try to convert the data content for a community descriptor term to follow the definition of one of the standard DwC terms. If a new community term is defined that could have been converted to one of the existing DwC terms, interoperability with biodiversity datasets from other communities will be broken.

(3) The next step is to create the new community's extension to the GBIF Vocabularies site (http://vocabularies.gbif.org, verified 31 Jan 2011), and to add the agreed community terms.

(4) When the registration and description of all terms for the new extension is ready at the Vocabularies site, the extension can be synchronized with the GBRDS (http://gbrds.gbif.org, verified 31 Jan 2011).

(5) Changing the workflow status first from Draft to Review will synchronize the new extension to the development GBRDS), and

(6) changing the workflow status from Review to Published will upload the final extension to the production version of the GBRDS (Harman, 2009). Please note that after synchronization with the official production version of the GBRDS, any modifications to the extension (however minor) need to be released with a new version number. The new community extension should now be ready for use with the GBIF IPT (http://code.google.com/p/gbif-providertoolkit/, verified 31 Jan 2011).

The GBIF Secretariat will provide support with the implementation of a new extension, if needed.

Standards are paramount

Since the genebank community already had established information standards, the development of a draft extension to the Darwin Core (DwC-germplasm) and the subsequent testing of the new prototype information publishing toolkit from GBIF (GBIF IPT), progressed so fast and with relatively few problems.

Efficient access to distributed PGR datasets stimulates novel uses

Research integrating genebank passport data (georeferenced occurrence data for the original collecting site) with phenotypic measurements (characterization and evaluation data, C&E) and with ecoclimatic layers has opened new possibilities for a rational utilization of genebank materials (Bhullar et al., 2009; Endresen, 2010) using the FIGS (Focused Identification of Germplasm Strategy) approach (Mackay and Street, 2004). The efficient application of the FIGS approach depends on the availability of germplasm passport and trait evaluation data.

The analysis of gaps in the genebank collections to guide the planning of rational germplasm collection expeditions to complement the genebank collections with novel and insufficiently sampled genetic diversity is also dependent on the availability of genebank passport data (Jarvis et al., 2003, 2005, 2009; Ramírez-Villegas et al., 2010). Such data analysis experiments to identify the ecological environment linked to a target trait property, or the genetic gaps of the genebank collections, will of course benefit from occurrence data on crop wild relatives provided from other communities. The value of external data from outside the genebank community, in such studies, strengthens the argument for the development of common semantic data standards (like the Darwin Core) and data exchange protocols (syntactic data standards like the GBIF IPT).

Limited access to genebank accession-level information is a bottleneck to the efficient use of genebank material (FAO, 2010), but also to the development of novel uses for the associated data. The authors of this manuscript propose the Darwin Core extension for genebanks and its implementation in the GBIF Integrated Publishing Toolkit (IPT) as a contribution for an upgrade of the current data exchange mechanism for genebank datasets.

Future work

After the first experiences with the deployment of the Darwin Core extension for genebanks, a useful next step will be to seek ratification of the extension as a TDWG standard. The genebank community has long and successful experience with the development and maintenance of descriptor standards, in particular through the work at Bioversity International (Bioversity International, 2007; Gotor et al., 2008). However, as discussed above, one of the major achievements with the DwC-germplasm is the interoperability with other biodiversity information standards and communities outside the genebank community. The ratification of genebank standards like the DwC-germplasm in TDWG will contribute to improved information interoperability.

The first version of the DwC-germplasm included the proposed EPGRIS3 descriptors for evaluation and characterization data. The sharing of trait datasets for germplasm has received renewed attention with the second report on the state of the world's plant genetic resources for food and agriculture (FAO, 2010). These descriptors need further work after the first experiences with the sharing of germplasm trait datasets.

The implementation during the last years of new international regulations for the sharing of benefits for the use of plant genetic resources prescribes the reporting of the distribution of seed samples (defined by the ITPGRFA, Annex 1). If the terms to describe and report these seed

distributions are developed and included to the DwC-germplasm, then the GBIF IPT could be used to report seed distributions to the Governing Body of the ITPGRFA.

CONCLUSIONS

The Darwin Core germplasm extension provides access to the GBIF bioinformatics infrastructure, including the GBIF Integrated Publishing Toolkit (IPT). Using the GBIF IPT and the Darwin Core germplasm extension, genebanks can now share PGR datasets with each other. This new data exchange mechanism will make the development of distributed PGR information networks easier. Implementation of general biodiversity information standards and toolkits will ensure the interoperability of genebank datasets with other biodiversity datasets.

ACKNOWLEDGEMENTS

Thanks for great help and support from colleagues of the GBIF secretariat, GBIF network participants, BioCASE team, DiGIR team, TDWG community, Bioversity International, Nordic Genetic Resource Center, EURISCO participants, EPGRIS3 participants, and the ECPGR Documentation and Information Network Coordinating Group. The mapping of MCPD to ABCD was carried out in collaboration with Walter Berendsohn and Javier de la Torre and their colleagues from the Botanical Garden and Botanical Museum Berlin-Dahlem, Germany. Kehan Harman provided valuable feedback regarding the http://vocabularies.gbif.org system. John Wieczorek and the other members of the Darwin Core task force provided advice and support during the development of the genebank extension to the Darwin Core. Tim Robertson, Markus Döring, José Cuadra and Samy Gaiji at the GBIF secretariat provided support and assistance with the deployment of the DwC-germplasm in the GBIF IPT software. The EURISCO secretariat and colleagues at Bioversity International including Sónia Dias, Milko Skofic, Elizabeth Arnaud and Michael Mackay provided great assistance with the development of the DwC-germplasm and in particular with the feasibility study to deploy the GBIF IPT for the first European genebanks.

REFERENCES

Alercia, A., S. Diulgheroff, T. Metz (2001). FAO/IPGRI Multi-crop passport descriptors, December 2001. International Plant Genetic Resources Institute (IPGRI) / Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. Available at http://apps3.fao.org/wiews/mcpd/MCPD_Dec2001_EN.pdf (verified 31 Jan 2011)

Anonymous (1983). ARS takes the reins on GRIN from LISA. Diversity 2(1): 6-7. Available at http://www.wlbcenter.org/journal/vol2_1.pdf (verified 31 Jan 2011).

Anonymous (1984). Germplasm Resources Information Network now available to scientific community. Diversity 2(3): 10. Available at http://www.wlbcenter.org/journal/vol2_3.pdf (verified 31 Jan 2011).

Bareš, I. (ed.) (1974). *Shirokiy unifitsirovannyy klassifikator SEV i Mezhdunarodnyy klassifikator SEV roda Triticum* [The International COMECON list of descriptors for the genus *Triticum* L.] Institut Genetiki i Selektsii, Praga-Ruzyne, Czech Republic. 128 p.

Berendsohn, W.G. (2002). BioCASE - A Biological Collection Access Service for Europe. Alliance News 29(6): 6-7.

Berendsohn, W.G. (ed.) (2005). ABCD Schema - Task group on access to biological collection data, [Online]. Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin, Germany. Available at http://www.bgbm.org/TDWG/CODATA/default.htm (verified 31 Jan 2011).

Berendsohn, W., and H. Knüpffer (2006). Draft mapping of EURISCO descriptors to ABCD 2.06, [Online]. Published online by the Berlin Botanical Garden (BGBM). Available at http://www.bgbm.org/tdwg/codata/schema/Mappings/EURISCO-2-ABCD.pdf (verified 31 Jan 2011).

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: 9519-9524. DOI: 10.1073/pnas.0904152106.

Biodiversity Information Standards (TDWG), [Online]. Available at http://www.tdwg.org (verified 31 Jan 2011).

Bioversity International (2007). Guidelines for the development of crop descriptor lists. Bioversity Technical Bulletin Series. Bioversity International, Rome, Italy. xii+72p. ISBN: 978-92-9043-792-1. Available at http://www.bioversityinternational.org/index.php?id=19&user_bioversitypublications_pi1[showUid]=3070 (verified 31 Jan 2011).

Bjarnason, S. (1989). *NGB's informationshandtering* [The NGB Information Management]. *In*: Nordiska Genbanken 1979 - 1989. NGB i samhällets tjänst. Möte 19 - 21 Juni, 1989, Hotel Sparta, Lund. NGB, Lund, 1989. [In Swedish]

Bjarnason, S., and M. Niklasson (compilers) (1989). The Nordic Barley Catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-00-5. ISSN: 1100-3456. NGB publications no 1. Available at http://www.ngb.se/Databases/Download/ (verified 31 Jan 2011).

Bruskiewich, R.M., A.B. Cosico, W. Eusebio, A.M. Portugal, L.M. Ramos, M.T. Reyes, M.A.B. Sallan, V.J.M. Ulat, X. Wang, K.L. McNally, R.S. Hamilton, and C.G. McLaren (2003). Linking genotype to phenotype: the International Rice Information System (IRIS). Bioinformatics 19(s1): i63-i65. DOI: 10.1093/bioinformatics/btg1006.

Bruskiewich, R., G. Davenport, T. Hazekamp, T. Metz, M. Ruiz, R. Simon, M. Takeya, J. Lee, M. Senger, G. McLaren, and T.v. Hintum (2006). The Generation Challenge Programme (GCP): Standards for crop data. OMICS, A Journal of Integrative Biology 10(2): 215-219. DOI: 10.1089/omi.2006.10.215.

Bruskiewich, R., M. Senger, G. Davenport, M. Ruiz, M. Rouard, T. Hazekamp, M. Takeya, K. Doi, K. Satoh, M. Costa, R. Simon, J. Balaji, A. Akintunde, R. Mauleon, S. Wanchana, T. Shah, M. Anacleto, A. Portugal, V.J. Ulat, S. Thongjuea, K. Braak, S. Ritter, A. Dereeper, M. Skofic, E. Rojas, N. Martins, G. Pappas, R. Alamban, R. Almodiel, L.H. Barboza, J. Detras, K. Manansala, M.J. Mendoza, J. Morales, B. Peralta, R. Valerio, Y. Zhang, S. Gregorio, J. Hermocilla, M. Echavez, J.M. Yap, A. Farmer, G. Schiltz, J. Lee, T. Casstevens, P. Jaiswal, A. Meintjes, M. Wilkinson, B. Good, J. Wagner, J. Morris, D. Marshall, A. Collins, S. Kikuchi, T. Metz, G. McLaren, and T.v. Hintum (2008). The Generation Challenge Programme platform: Semantic standards and workbench for crop science. International Journal of Plant Genomics, 2008: 1-7. DOI: 10.1155/2008/369601.

CBD (1993). Convention on Biological Diversity. Earth summit 5 June 1992, Rio de Janeiro, Brazil. United Nations, Treaty Series, Vol. 1760, I-30619. Available online at http://www.cbd.int/convention/convention.shtml (verified 31 Jan 2011).

CWRIS, PGR Forum Crop Wild Relative Information System [Online, Software]. Available at http://www.pgrforum.org/cwris/ (verified 31 Jan 2011).

CWRML, PGR Forum Crop Wild Relative Markup Language [Online]. Available at http://pgrforum.org/CWRML.htm (verified 31 Jan 2011).

Darwin Core Task Group (2009). Darwin Core [Online]. Available at http://rs.tdwg.org/dwc (verified 31 Jan 2011). Biodiversity Information Standards (TDWG; www.tdwg.org).

DeLacy, I.H., P.N. Fox, G. McLaren, R. Trethowan, and J.W. White (2009). A conceptual model for describing processes of crop improvement in database structures. Crop Science 49: 2100-2112. DOI: 10.2135/cropsci2009.01.0020.

Dublin Core Metadata Initiative (DCMI, [Online]. Available at http://dublincore.org/ (verified 31 Jan 2011).

Dragavtsev, V., L. Gorbatenko, L. Bagmet, V. Funtova (compilers) (1999). Delectus Seminum, 1999-2004. The N.I. Vavilov All-Russian Scientific Research Institute of Plant Industry (VIR), St. Petersburg, Russia.

DwC-germplasm, Darwin Core extension for genebanks, [Online]. Available at http://rs.nordgen.org/dwc/, and at http://code.google.com/p/darwincore-germplasm/ (verified 31 Jan 2011).

ECP/GR (1997) European Central Crop Databases (ECCDB) On-line databases training workshop, 8-10 June 1997, Bonn, Germany, [Online]. Available at http://www.ecpgr.cgiar.org/eccdb1/bonn2.htm (verified 31 Jan 2011).

ECPGR Barley Database. Hosted by IPK Gatersleben, Germany, [Online]. Available at http://barley.ipk-gatersleben.de/genres/ (verified 31 Jan 2011).

ECPGR *Phleum* Database. Hosted by the Nordic Genetic Resources Center. Available at http://www.nordgen.org/ecpgr/ (verified 31 Jan 2011).

Ellerström, S. (1982). Some notes on the origin of the Nordic Gene Bank. Journal of the Swedish Seed Association 92: 87-91.

Endresen, D.T.F. (2010). Predictive association between trait data and ecogeographic data for Nordic Barley Landraces Crop Science 50 (6) DOI: 10.2135/cropsci2010.03.0174.

Endresen, D.T.F., J. Bäckman, H. Knüpffer, S. Gaiji (2006). Exchange of germplasm datasets with PyWrapper/BioCASE. p. 8. *In*: Belbin, L., A. Rissoné, A. Weitzman (eds). Proceedings of TDWG 2006. Taxonomic Databases Working Group, St. Louis, MI, USA. Available at http://www.tdwg.org/proceedings/article/view/64 (verified 31 Jan 2011).

Endresen, D.T.F., S. Gaiji, T. Robertson (2009). DarwinCore Germplasm extension and deployment in the GBIF infrastructure. p. 78. *In*: Weitzman, A.L. (ed). Proceedings of TDWG 2009, Taxonomic Databases Working Group, Montpellier, France. Available at http://www.tdwg.org/proceedings/article/view/464 (verified 31 Jan 2011).

Endresen, D.T.F., V., Kukk, R. Baltrenas (2005a). Regional Nordic-Baltic database cooperation. Nordic Gene Resources, livestock, crops, forest trees, volume 4 (4): 15.

Endresen, D.T.F., B. Skovmand, J. Bäckman (2005b). Integrated generic regional genetic resources information system. ASA-CSSA-SSSA International Annual Meetings. http://crops.confex.com/crops/2005am/techprogram/P6077.HTM (verified 31 Jan 2011).

EPGRIS3, Establishment of a European Plant Genetic Resources Information Infra-Structure, phase 3, [Online]. Available at http://www.epgris3.eu/ (verified 31 Jan 2011).

EPGRIS3 wiki [Online] Available at http://www.nordgen.org/epgris3/wiki/index.php/DwC_Germplasm (verified 31 Jan 2011).

EURISCO. EURISCO Catalogue of *ex situ* plant genetic resources in Europe [Online database]. Available at http://eurisco.ecpgr.org/ (verified 31 Jan 2011)

EURISCO (2002). EURISCO uploading mechanism - Technical notes, Draft July 4, 2002. Retrieved from http://eurisco.ecpgr.org/about/documents/epgris_uploading_mechanism.pdf (verified 31 Jan 2011).

EURISCO (2003). EPGRIS final meeting. 11-13 September 2003, Prague, Czech Republic. PGR Documentation and Information in Europe. Towards a sustainable and user-oriented information infrastructure. Conference report available at http://www.ecpgr.cgiar.org/Networks/Info_doc/FinalMeetingReports.htm (verified 31 Jan 2011).

EURISCO (2011). Descriptors for uploading information from National Inventories to EURISCO. http://eurisco.ecpgr.org/documents/MCPD%20_EURISCO_Descriptors_111.pdf (verified 31 Jan 2011). [Online document marked as last updated 11 Jan 2011]

FAO (1996). Global Plan of Action for the Conservation and Sustainable Utilization of Plant Genetic Resources for Food and Agriculture and the Leipzig Declaration adopted by the International Technical Conference on Plant Genetic Resources, Leipzig, Germany 17-23 June 1996. Food and Agriculture Organization of the United Nations, Rome, Italy. Available at http://www.fao.org/agriculture/crops/core-themes/theme/seeds-pgr/gpa/en/ (verified 31 Jan 2011).

FAO (2002). International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA). Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. Available at http://www.planttreaty.org (verified 31 Jan 2011).

FAO (2009). International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA). Food and Agriculture Organization of the United Nations, Rome, Italy. Second edition. Available at http://www.planttreaty.org (verified 31 Jan 2011).

FAO (2010). The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Commission on Genetic Resources for Food and Agriculture (CGRFA), Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. ISBN 978-92-5-106534-1. Available at http://www.fao.org/docrep/013/i1500e/i1500e00.htm (verified 31 Jan 2011).

FAO ITPGRFA (2008). Information management in support of the global system for the conservation and sustainable use of plant genetic resources for food and agriculture - Global Information on Germplasm (GIG). Report from the second technical consultation on information technology support for the implementation of the Multilateral System of Access and Benefit Sharing of the International Treaty, Rome, 2-3 December 2008. FAO, Rome, Italy. IT/GB-2/07/REPORT. (Principal investigators: Michael Mackay and Elizabeth Arnaud). Available at ftp://ftp.fao.org/ag/agp/planttreaty/gb3/tcit2/tcit2i02.pdf (verified 31 Jan 2011).

Finlay, K.W, and F. Konzak (1970). Information storage and retrieval. p. 461-465. *In*: Frankel, O.H. and E. Bennett (eds). Genetic resources in plants - Their exploration and conservation. IBP Handbook No 11. International Biological Programme, London, UK / Blackwell Scientific Publications, Bell and Bain Ltd., Glasgow, UK.

Fox, P.N., C. Lopez, B. Skovmand, H. Sanchez, R. Herrera, J.W. White, E. Duveiller, and M. van Ginkel (1996). International Wheat Information System (IWIS), Version 1 [CD-ROM]. CIMMYT, Mexico.

Fox, P.N., and B. Skovmand (1996). The International Crop Information System (ICIS) - Connects genebank to breeder to farmer's field. p. 317–326. *In*: Cooper, M. and G.L. Hammer (ed). Plant adaptation and crop improvement. CAB Int., Wallingford, UK. 636 p. ISBN: 9780851991085.

Gaiji, S., S. Dias, D.T.F. Endresen, and T. Franco (2008). *Desarrollo de un sistema global de información a nivel de accesiones en apoyo al Tratado Internacional sobre los Recursos Fitogenéticos para la Alimentación y la Agricultura* [Building a global accession level information system in support of the International Treaty on Plant Genetic Resources for Food and Agriculture - ways forward in the Americas]. *Recursos Naturales y Ambiente* [Journal of Natural Resources and the Environment] 53: 126-135. Available at http://web.catie.ac.cr/informacion/RFCA/rev53/rna53_p126_135.pdf (verified 31 Jan 2011). [In Spanish with English summary]

Gaiji, S., D.T.F. Endresen, J. Nordling, S. Dias, and E. Arnaud (2010). Beyond Darwin Core: Challenges in mobilizing richer content. p. 15-16. *In:* Proceedings of TDWG (2010), Woods Hole Massachusetts, USA. Available at http://www.tdwg.org/fileadmin/2010conference/documents/Provisional_Proceedings_of_TDWG_2010.pdf (verified 31 Jan 2011).

Global Biodiversity Information Facility (GBIF) [Online]. Available at http://www.gbif.org (verified 31 Jan 2011).

GBIF (2001). Executive Summary of the 1st Meeting of the Governing Board of the Global Biodiversity Information Facility (GBIF). Available at http://www2.gbif.org/GB1_ExecutiveSummary.pdf (verified 31 Jan 2011).

GBIF Integrated Publishing Toolkit (IPT) [Software]. Available at http://code.google.com/p/gbif-providertoolkit/ (verified 31 Jan 2011).

GBIF. Global Biodiversity Resources Discovery System (GBRDS) [Online]. Available at http://gbrds.gbif.org/, and at http://code.google.com/p/gbif-registry/ (verified 31 Jan 2011).

GBIF Vocabularies [Online]. Available at http://vocabularies.gbif.org (verified 31 Jan 2011).

GCP Passport XML schema. Generation Challenge Program (GCP) Subprogramme 4, Bioinformatics [Online]. Available at http://gcpcr.grinfo.net/include/webservices/schema-documentation.php (verified 31 Jan 2011).

Generation Challenge Program Central Registry (GCP CR) [Online]. Available at http://gcpcr.grinfo.net (verified 31 Jan 2011), and for download at http://cropforge.org/projects/gcpcr/ (verified 31 Jan 2011).

Gotor, E., A. Alercia, V. Ramanatha Rao, J. Watts, and F. Caracciolo (2008). The scientific information activity of Bioversity International: the descriptor lists. Genetic Resources and Crop Evolution 55(5): 757-772. DOI: 10.1007/s10722-008-9342-x.

Hammer, K. (2003). R. Mansfeld's scientific influence on genetic resources research. p. 7-16. *In*: Knüpffer, H. & J. Ochsmann (eds). Rudolf Mansfeld and Plant Genetic Resources. Proceedings of a symposium dedicated to the 100th birthday of Rudolf Mansfeld, Gatersleben, Germany, 8-9 October 2001. Schriften zu Genetischen Ressourcen. Band 22. Zentralstelle für Agrardokumentation und -information (ZADI), Bonn, Germany. ISSN: 0948-8332. Available at http://www.genres.de/fileadmin/SITE_GENRES/downloads/schriftenreihe/Band22_Gesamt.pdf (verified 31 Jan 2011).

Harman, K. (2009). GBIF Vocabulary and Extension Server, User Documentation [Online]. Global Biodiversity Information Facility, Copenhagen, Denmark. Available at http://vocabularies.gbif.org/sites/vocabularies.gbif.org/files/UserDocumentation.pdf (verified 31 Jan 2011).

Harrer, S., and A. Omelchenko (1998). VIR information system updated. Plant Genetic Resources Newsletter 13: 5. Available at http://www.bioversityinternational.org/nc/publications/publication/issue/newsletter_for_europe-17.html (verified 31 Jan 2011).

Hazekamp, T., J. Serwiński, and A. Alercia (1997). Appendix II. Multicrop passport descriptors (final version). p. 97-90. *In*: Lipman, E., M.W.M. Jongen, Th.J.L. van Hintum, T. Grass, and L. Maggioni (eds). Central crop databases: Tools for plant genetic resources management. International Plant Genetic Resources Institute (IPGRI), Rome, Italy/CGN, Wageningen, Netherlands. ISBN 92-9043-320-5.

Heywood, V.H. (1964). Some aspects of seed lists and taxonomy. Taxon 13(3): 90-94. Available at http://www.jstor.org/pss/1216623 (verified 31 Jan 2011).

Huldén, M. (1997). European *Phleum* Database. IPGRI Newsletter for Europe 10: 9. Available at http://www.bioversityinternational.org/nc/publications/publication/issue/newsletter_for_europe-26.html (verified 31 Jan 2011).

Huldén, M. (1999). Development of the information system at the NGB 1989-1999. p. 26-28. *In*: Nordic Gene Bank 1979 - 1999. Nordic Gene Bank, Alnarp, Sweden. ISSN 1100-3456, ISRN NGB-S--35--SE.

Huldén, M., B. Lund, G.P. Poulsen, E. Thörn, and J. Weibull (1998). The Nordic commitment: Regional and international collaboration on plant genetic resources. Plant Varieties and Seeds 11: 1-13.

IBPGR (1976). First report of the advisory committee on the genetic resources communication, information and documentation system (GR/CIDS). International Board for Plant Genetic Resources, Rome, Italy. AGPE:IBPGR/76/7.

IBPGR (1977). Descriptors for the cultivated potato and for the maintenance and distribution of germplasm collections. International Board for Plant Genetic Resources (IBPGR), Rome, Italy. 47 p.

IBPGR (1978). Descriptors for wheat and *Aegilops*: a minimum list. International Board for Plant Genetic Resources (IBPGR), Rome, Italy. 25 p.

International Crop Information System (ICIS), Ver. 1.0. [Software]. International Rice Research Institute (IRRI), Manila, Philippines. Available online at http://www.icis.cgiar.org/icis/ (verified 31 Jan 2011).

International Rice Information System (IRIS) [Online database]. International Rice Research Institute (IRRI), Manila, Philippines. Available online at http://beta.irri.org/seeds/ (verified 31 Jan 2011).

IPGRI (2001). European PGR Information Infra-Structure. IPGRI Newsletter for Europe 22: 5. Available at http://www.bioversityinternational.org/nc/publications/publication/issue/newsletter_for_europe-13.html (verified 31 Jan 2011).

IPGRI (2002). EPGRIS entering its third and final year. EURISCO and central crop databases. IPGRI Newsletter for Europe 25: 5. Available at

http://www.bioversityinternational.org/fileadmin/bioversity/publications/pdfs/822.pdf (verified 31 Jan 2011).

IPGRI (2003). Final EPGRIS conference and ECP/GR documentation and information network meeting. Newsletter for Europe 27: 4. Available at http://www.bioversityinternational.org/nc/publications/publication/issue/newsletter_for_europe-3.html (verified 31 Jan 2011).

Jarvis, A., M.E. Ferguson, D.E. Williams, L. Guarino, P.G. Jones, H.T. Stalker, J.F.M. Valls, R.N. Pittman, C.E. Simpson, and P. Bramel (2003). Biogeography of wild *Arachis*: Assessing conservation status and setting future priorities. Crop Science 43:1100–1108. DOI: 10.2135/cropsci2003.1100.

Jarvis, A., J. Ramirez, N. Castañeda, S. Gaiji, L. Guarino, H. Tobón, D. Amariles (2009). Value of a coordinate: geographic analysis of agricultural biodiversity. p. 6-7 *In*: Weitzman, A.L. (ed). Proceedings of TDWG 2009, Taxonomic Databases Working Group, Montpellier, France. Available at http://www.tdwg.org/proceedings/article/view/555 (verified 31 Jan 2011).

Jarvis, A., K. Williams, D. Williams, L. Guarino, P.J. Caballero, and G. Mottram (2005). Use of GIS for optimizing a collecting mission for rare wild pepper (*Capsicum flexuosum* Sendtn.) in Paraguay. Genetic Resources Crop Evolution 52:671-682. DOI: 10.1007/s10722-003-6020-x.

Jõgeva Plant Breeding Institute (Jõgeva PBI) (2005). Report of activities of 2005. Available at http://www.sordiaretus.ee/files/GP/Aruanded_2005/JPBIeng.pdf (verified 31 Jan 2011).

Knüpffer, H. (1983). Computer in Genbanken - eine Übersicht. Kulturpflanze 31: 77-143. DOI: 10.1007/BF02000699.

Knüpffer, H. (compiler) (1987). European Barley List. Vol. 1: Introduction. 82 p. Vol. 2, Part 1: Cultivars, lines and special resources. Part 2: Collected material, unnamed accessions. Part 3: Wild species, species hybrids. 829 p. Zentralinstitut für Genetik und Kulturpflanzenforschung Gatersleben; International Board for Plant Genetic Resources, Rome.

Knüpffer, H. (1988) The European Barley Database of the ECP/GR: An introduction. Kulturpflanze 36: 135-162. DOI: 10.1007/BF01986957.

Knüpffer, H. (1995). Central crop databases. p. 51-62. *In*: Hintum, Th.J.L. van, M.W.M. Jongen, and T. Hazekamp (eds). Standardization in plant genetic resources documentation. Centre for Genetic Resources, The Netherlands (CGN), Wageningen, The Netherlands.

Knüpffer, H. (1997). Options and approaches to providing on-line access to databases. Part III: FoxPro on-line databases on an institution server. Report, "European Central Crop Databases (ECCDB) - On-line Databases Training Workshop", Bonn, 8-10 June 1997. Internet Publication under the ECP/GR Platform (http://www.ecpgr.cgiar.org/eccdb1/, verified 31 Jan 2011) as http://www.ecpgr.cgiar.org/eccdb1/Fox-serv.htm (verified 31 Jan 2011).

Knüpffer, H. (ed.) (1999a). Index Seminum quae pro mutua commutatione offert Institut für Pflanzengenetik und Kulturpflanzenforschung Gatersleben 2000. IPK, Gatersleben, 131 pp.

Knüpffer, H. (ed.) (1999b). Supplementum Cultivarorum ad Index Seminum Gaterslebensis 2000. IPK, Gatersleben, 282 pp.

Knüpffer, H., N. Biermann, D.T. Endresen, P. Kolasinski, W. Podyma, and J. de la Torre (2004). Genebanks as GBIF data providers – first experiences. *In*: Proceedings of TDWG 2004. Available at http://www.nhm.ac.uk/hosted_sites/tdwg/2004meet/TDWG_2004_AbstractsPapers.htm (verified 31 Jan 2011).

Knüpffer, H., D.T.F. Endresen, I. Faberová, S. Gaiji (in press). Integrating genebanks into biodiversity information networks. *In*: Proceedings, 18th EUCARPIA Genetic Resources Section Meeting: Plant Genetic Resources and their Exploitation in the Plant Breeding for Food and Agriculture, Piešťany, Slovak Republic, May 23-26, 2007.

Knüpffer, H., D.T.F. Endresen, S. Gaiji (2007). Integrating genebanks into biodiversity information networks. p. 34-35. *In*: Book of Abstracts, 18th EUCARPIA Genetic Resources Section Meeting: Plant Genetic Resources and their Exploitation in the Plant Breeding for Food and Agriculture, May 23-26, 2007. Available at http://www.eucarpia.org/03publications (verified 31 Jan 2011).

Konzak, C.F., and B. Sigurbjörnsson (1966). International cooperation in standardization of procedures in crop research data recording. Fifth Yugoslav Symposium on Research in Wheat. Contemporary Agriculture 11-12, pp. 691-696.

Kurki, A. (ed.) (1986). Apricot (*Prunus armenica*) from ECP/GR prunus central data base, July 1986. European Cooperative Programme for the conservation and exchange of crop Genetic Resources (ECP/GR) and Nordic Gene Bank (NGB), Alnarp, Sweden.

Lipman, E., M.W.M. Jongen, Th.J.L. van Hintum, T. Grass, and L. Maggioni (eds.) (1997). Central crop databases: Tools for plant genetic resources management. International Plant Genetic Resources Institute, Rome, Italy/CGN, Wageningen, Netherlands. ISBN 92-9043-320-5.

Lindeberg, G., K. Sandvad, and S. Blixt (1981). *PM rörande Nordiska genbankens databehandlingsstruktur och dess utbygnad samt uppdateringsbehov av nuvarande ADB-anläggning* [Report on requirements for the upgrade of the information infrastructure at the Nordic Gene Bank]. Nordiska Genbanken, Lund. (Retrieved from the NordGen correspondence archive, cornum 8100154 and cornum 8100030) [In Swedish]

Loskutov, I.G. (1999). Vavilov and his institute. A history of the world collection of plant genetic resources in Russia. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. ISBN: 92-9043-412-0.

Loskutov, I.G., and E.E. Ryabchenko (compilers) (2002). Catalogue of the VIR World Collection, edition 735, Oats. The N.I. Vavilov All-Russian Scientific Research Institute of Plant Industry, St. Petersburg, Russia. [In Russian].

Mackay, M., D.T.F. Endresen, S. Dias, R. Sood, K. Viparthi, M. Skofic, A. Alercia, T. Franco, F. Atieno, X. Scheldeman, A. Shamsie, S. Louafi, P. Cyr E. Arnaud (2008). A global information system for the conservation and sustainable use of plant genetic resources for food and agriculture (PGRFA). p. 78. *In*: Weitzman, A.L., and L. Belbin (eds). Proceedings of TDWG 2008, Fremantle, Australia. Available at http://www.tdwg.org/proceedings/article/view/396 (verified 31 Jan 2011).

Mackay, M.C., and K. Street (2004). Focused Identification of Germplasm Strategy - FIGS. Cereals. p. 138-141. *In*: Black, C.K., J.F. Panozzo, and G.J. Rebetzke (eds). Proceedings of the Australian Cereal Chemistry Conf., 54th, and the Wheat Breeders' Assembly, 11th, Canberra, ACT. 21-24 September 2004. Royal Australian Chemical Institute, Melbourne, Australia.

Maggioni, L. (ed.) (2005). Summary of a Network Coordinating Group on Documentation and Information and the EURISCO Advisory Group. International Plant Genetic Resources Institute (IPGRI), Rome, Italy. Available at http://www.bioversityinternational.org/fileadmin/bioversity/publications/pdfs/1051.pdf (verified 31 Jan 2011).

Maggioni, L. (ed.) (2007). Minutes of a joint meeting of the documentation and information network coordinating group and the EURISCO advisory group. Planning for the continuation of EPGRIS, 2-3 April 2007. Bioversity International, Rome, Italy. Available at http://www.ecpgr.cgiar.org/Networks/Info_doc/DINCG_AdvGR_ROMEApril07.pdf (verified 31 Jan 2011).

Maggioni, L. (2009). C&E data in ECCDBs. Presentation given at the EPGRIS3 planning meeting on 7 May 2009 in Bonn, Germany. 13 p. Available at http://www.epgris3.eu/EPGRIS3 20090507workshop.htm, http://www.epgris3.eu/docs/activities/2-05/LM - C&E in ECCDBs.pdf (verified 31 Jan 2011).

Maggioni, L. (ed.) (2010). Report of the ECPGR documentation and information network coordinating group, forth meeting, 17-18 February 2010, Maccarese, Rome, Italy. Bioversity International, Rome, Italy. 22 p. Available at http://www.ecpgr.cgiar.org/Networks/Info_doc/Doc&Info NCG Fourth Meeting final for Web 260510.pdf (verified 31 Jan 2011).

McLaren, C.G., R.M. Bruskiewich, A.M. Portugal, and A.B. Cosico (2005). The International Rice Information System. A platform for Meta-analysis of rice crop data. Plant Physiology 139: 637-642. DOI: 10.1104/pp.105.063438.

Moore, J.D., S.P. Kell, J.M. Iriondo, B.V. Ford-Loyd, and N. Maxted (2008). CWRML: Representing crop wild relative conservation and use data in XML. BMC Bioinformatics 9: 116. DOI: 10.1186/1471-2105-9-116. Available at http://www.biomedcentral.com/1471-2105/9/116/pdf and at http://pgrforum.org/CWRML.htm (verified 31 Jan 2011).

NGB (1989). Nordic Gene Bank seed catalogue 1989. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-06-4. ISSN: 110-3456. NGB publications no 7.

NGB (1991). Nordiska Genebanken 1979 - 1989. Nordic Gene Bank, Alnarp, Sweden. ISSN: 1100-3456. ISRN: NGB-S--12--SE. NGB publications no 12. [10-year jubilee, with summary from the first 10 years]

Niklasson, M, and S. Bjarnason (compilers) (1989a). The European cherry catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-01-3. ISSN: 1100-3456. NGB publications no 2.

Niklasson, M, and S. Bjarnason (compilers) (1989b). The European apricot catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-02-1. ISSN: 1100-3456. NGB publications no 3.

Niklasson, M. (compiler) (1989a). The European almond catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-03-X. ISSN: 1100-3456. NGB publications no 4.

Niklasson, M. (compiler) (1989b). The European peach catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-04-8. ISSN: 1100-3456. NGB publications no 5.

Niklasson, M. (compiler) (1989c). The European plum catalogue. Nordic Gene Bank, Alnarp, Sweden. ISBN: 91-87814-05-6. ISSN: 1100-3456. NGB publications no 6.

Omelchenko, A., S. Alexanian, S. Harrer, and A. Serbin (2003). The information system on plant genetic resources of the N.I. Vavilov All-Russian Institute of Plant Industry (VIR). p. 298-300. *In*: Knüpffer, H., and J. Ochsmann (ed). Rudolf Mansfeld and Plant Genetic Resources. Proceedings of a symposium dedicated to the 100th birthday of Rudolf Mansfeld, Gatersleben, Germany, 8-9 October 2001. Schriften zu Genetischen Ressourcen. Band 22. Zentralstelle für Agrardokumentation und -information (ZADI), Bonn, Germany. ISSN: 0948-8332. Available at http://www.genres.de/fileadmin/SITE_GENRES/downloads/schriftenreihe/Band22_Gesamt.pdf (verified 31 Jan 2011).

Palmstierna, H., G. Julen, L. Kåhre, and S.O. Myresten (1975). *Genbank för jordbruk och trädgårdsnäring: Betänkande avgivet av Genbankutredningen* [Genebank for Agriculture and Horticulture: report from the Genebank Committee]. Jordbruksdepartementet, Liber Förlag, Stockholm, Sweden. 78 p. ISSN: 0346-5667, ISBN: 9138023539, Ds Jo 1975:5. [In Swedish with English summary]

Podyma, W. (2001) European *Secale* Database. p. 30-31. *In*: Maggioni, L., and O. Spellman (compilers). Report of a network coordinating group on cereals, *ad hoc* meeting, 7-8 July 2000, Radzików, Poland. International Plant Genetic Resources Institute (IPGRI), Rome, Italy.

Postman, J., K. Hummer, T. Ayala-Silva, P. Bretting, T. Franko, G. Kinard, M. Bohning, G. Emberland, Q. Sinnott, M. Mackay, P. Cyr, M. Millard, C. Gardner, L. Guarino, and B. Weaver (2010). GRIN-Global: An international project to develop a global plant genebank information management system. Acta Horticulturae 859: 49-58. Available at http://ddr.nal.usda.gov/handle/10113/43806 (verified 31 Jan 2011).

Ramírez-Villegas, J., C. Khoury, A. Jarvis, D.G. Debouck, and L. Guarino (2010). A gap analysis methodology for collecting crop genepools: A case study with *Phaseolus* beans. PLoS ONE 5(10): e13497. DOI: 10.1371/journal.pone.0013497.

Regel R.E. (1915). *Organizatsiya i deyatel'nost' Byuro po prikladnoy Botanike za pervoe dvadtsatiletie ego sushchestvovaniya* [Organization and activity of the Bureau of Applied Botany for the first twenty years of its existence]. Bulletin of the Bureau of Applied Botany 8(4/5): 327-767. (In Russian).

Richards, C.M., and G.M. Volk (2010). New challenges for data management in genebanks. Acta Horticulturae 859: 333-336. Available at http://ddr.nal.usda.gov/handle/10113/43816 (verified 31 Jan 2011).

Rogalewicz, V. (ed.), in collaboration with H. Knüpffer, V.A. Korneychuk, I.G. Lozanov, D.B. Plotnikov, J. Serwiński and I.A. Shvytov (1988). *Pasportnye deskriptory mezhdunarodnoy bazy dannykh geneticheskikh resursov stran-chlenov SEV* [Passport descriptors of the COMECON International Database of Genetic Resources]. Výzkumný ústav rostlinné výroby, Praha-Ruzyně, Czechoslovakia. 26 p.

Rydström, G. (1989). BIRS, biological information retrieval system user's guide. Nordic Gene Bank, Alnarp, Sweden.

Serwiński, J., and J. Konopka (1984). European catalogue of genus *Secale* L. First edition. European Cooperative Programme for the Conservation and Exchange of Crop Genetic Resources, International Board for Plant Genetic Resources (IBPGR), Rome, Italy.

SGRP (2011). Review of the FAO/IPGRI list of multi-crop passport descriptors (MCPD) [Online Blog]. Crop Genebank Knowledge Base. System-wide Genetic Resources Programme (SGRP), Consultative Group on International Agricultural Research (CGIAR). Available at http://cropgenebank.wordpress.com/2011/01/24/review-of-faoipgri-list-of-multi-crop-passport-descriptors-mcpd/ (verified 31 Jan 2011).

Stafleu, F.A. (1969). Botanic gardens before 1818. Boissiera 14: 31–46.

Stearn, S. (1998). From simple queries to complete data management: pcGRIN provides the answers. Diversity 14(3&4): 28. Available at http://www.wlbcenter.org/journal/vol14_2.pdf (verified 31 Jan 2011).

Stearn, W.T. (1971). Sources of information about botanic gardens and herbaria. Biological Journal of the Linnaean Society 3(3): 225-233. DOI: 10.1111/j.1095-8312.1971.tb00184.x.

Stein, B.R., and J. Wieczorek (2004). Mammals of the world: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics 1: 14-22.

Thompson, P.A. (1970). Seed Banks as a means of improving the quality of Seed Lists. Taxon 19(1): 59-62.

Thörn, E. (2006). SEEDNet update. IPGRI Newsletter for Europe 33:7. Available at http://www.bioversityinternational.org/nc/publications/publication/issue/newsletter_for_europe-6.html (verified 31 Jan 2011).

Townsend Peterson, A., D.A. Vieglais, A.G. Navarro Sigüenza, and M. Silvia (2003). A global distributed biodiversity information network: building the world museum. Bulletin of the British Ornithologists' Club 123A: 186-196.

USDA GRIN. The pcGRIN user manual, [Online]. USDA NPGS GRIN, Beltsville, Maryland, USA. Available at http://www.ars-grin.gov/npgs/pcgrin/manual/index.html (verified 31 Jan 2011).

Vieglais D.A., D.R.B Stockwell, C.M. Cundari, J. Beach, A.T. Peterson, and L. Krishtalka (1998). The species analyst: Tools enabling a comprehensive distributed biodiversity network. p. 144-147. *In*: Biodiversity, Biotechnology & Biobusiness, 2nd Asia Pacific Conference on Biotechnology, 23-27 November, Perth, Western Australia. Available at http://www.aseanbiodiversity.info/Abstract/52000571.pdf (verified 31 Jan 2011).

Weibel, S., J. Godby, E. Miller, and R. Daniel (1995). OCLC/NCSA Metadata workshop report. Dublin Core Metadata Initiative (DMCI). Available at http://dublincore.org/workshops/dc1/report.shtml (verified 31 Jan 2011).

Yndgaard, F. (1981). NGB's information system. *In*: Nordic Gene Bank concept meeting. Nordic Gene Bank, Lund. [Internal document, Nordic Gene Bank]

Yndgaard, F. and E. Kjellqvist (1982). *Orientering om Nordisk Genbank* [Introducing the Nordic Gene Bank]. Journal of the Swedish Seed Association 92: 75-86. [In Swedish with summary in English]

Paper II

Endresen, Dag Terje Filip (2010). Predictive Association between Trait Data and Ecogeographic Data for Barley Landraces. Crop Science 50(6): 2418-2430. DOI: 10.2135/cropsci2010.03.0174.

Available as open access at https://www.crops.org/publications/cs/articles/50/6/2418.

RESEARCH

Predictive Association between Trait Data and Ecogeographic Data for Nordic Barley Landraces

Dag Terje Filip Endresen*

ABSTRACT

Focused Identi cation of Germplasm (FIGS) is a new method to select plant genetic resources for the improvement of food crops. Traditional cultivars (landraces) and crop wild relatives (CWR) provide a valuable source for novel alleles in crop improvement programs, but conserved landraces and CWR often lack important documentation. Genebank collections worldwide provide ready access to plant genetic resources including online documentation. However, incomplete documentation, and in particular the lack of relevant characterization and evaluation data (traits), often limit the ef - cient use of plant genetic resources. This current study demonstrates how trait mining with the new FIGS method can be used to predict missing trait information for landraces. Ecogeographic data from the location of origin for 14 Nordic landraces of barley (Hordeum vulgare L.) was successfully correlated to morphological traits using a modern multilinear data modeling method (multilinear partial least squares [N-PLS]). This result suggests that trait mining can ef ciently be used as a targeted germplasm selection method and complement or replace the current core selection method in situations when the requirements for the trait mining method are ful lled.

Nordic Genetic Resources Center (NordGen), Smedjevägen 3, 230 53 Alnarp, Sweden. Received 24 Mar. 2010. *Corresponding author (dag.endresen@nordgen.org).

Abbreviations: FIGS, focused identi cation of germplasm strategy; G×E, genotype and environment interaction; N-PLS, multilinear partial least squares; PARAFAC, parallel factor analysis; PCA, principle component analysis; PLS, partial least squares.

To the development of the focused identi cation of germplasm strategy (FIGS) methodology (Mackay and Street, 2004). The overall goal of this new trait mining strategy for selection of germplasm is to improve access to relevant information and thus the usability of plant genetic resources conserved by ex situ gene banks worldwide. The application of multilinear data analysis and multiway methods for ecogeographic data on plant genetic resources is introduced here for the rst time. The target of this current study is to calibrate a computer simulation model to predict morphological traits in Nordic barley landraces. Ecogeographic data from the landrace location of origin is used to calibrate the computer model. The novel trait mining strategy is here demonstrated and developed further as a method for targeted selection of genebank accessions. Trait mining is presented as a complement and an alternative to the established core subset selection strategy (Frankel and Brown, 1984). In this study, modern multilinear data analysis methods (multiway data arrays, parallel factor analysis [PARAFAC], and N-PLS regression) are used to calibrate a computer model for the prediction of traits in Nordic barley landraces. The goal of the model calibration is to predict the unknown trait expression from the environmental pro le for a number of crop landraces. The ultimate goal is to rank a large

Published in Crop Sci. 50:2418–2430 (2010).
doi: 10.2135/cropsci2010.03.0174
Freely available online through the author-supported open-access option.
Published online 27 Sept. 2010.
© Crop Science Society of America | 5585 Guilford Rd., Madison, WI 53711 USA

All rights reserved. No part of this periodical may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or any information storage and retrieval system, without permission in writing from the publisher. Permission for printing and for reprinting the material contained herein has been obtained by the publisher.

number of crop landraces (conserved by ex situ crop genebanks) based on the predicted trait expression, to extract a reduced subset of samples. The final number of samples in the subset is chosen depending on the capacity for the planned field trial experiment.

Genetic Bottleneck during Domestication

As described, for example, by Tanksley and McCouch (1997), potentially economically valuable alleles were lost during the early crop domestication process. It is likely that a similar loss of some potentially economically valuable alleles also occurred during the cultivation process from the early primitive cultivars and landraces to modern cultivars (Gepts, 2006). Van de Wouw et al. (2010) argue that it was the switch from landraces to modern cultivars that caused most of this reduction in crop genetic diversity, and that crop genetic diversity in modern cultivars seems to be stable. Kolodinska Brantestam et al. (2003, 2004, 2007) reached similar conclusions for Nordic crop genetic diversity in barley. This study focuses on the improved access and utilization for traditional cultivars (landraces) as a source of novel alleles for crop improvement and future food security.

Core Selection Strategy

The increasing size of ex situ collections of plant genetic resources and the limited funds available for activities to develop characterization and evaluation data was early identified as a limitation to the use of germplasm collections (Frankel and Brown, 1984). The most common approach to identify traits for germplasm in crop genetic resources collections is to screen the available genebank samples in a field trial (Frankel, 1989; Frankel and Bennett, 1970). Most often the available relevant germplasm samples are much more abundant than the capacity of the field trial experiment. The limitations include both the human capacity to perform the field observations (manpower), as well as the limited size of the experimental field (land area). A typical example could be a plant breeder with the capacity to screen a few hundred germplasm samples, but presented with several hundred thousand relevant genebank germplasm accessions (Bouhssini et al., 2009). A common solution is to calculate a reduced set or a so-called "core selection" (Frankel, 1984). The core selection method aims to obtain a subset with minimum similarity between the entries to preserve the widest possible genetic diversity (van Hintum et al., 2000). The purpose of the core selection is thus to reduce the number of genebank accessions to a more manageable number. Brown (1989) proposed that the core selection could also be used as a method to "guide the breeder to other better sources" found outside the core collection, in the so-called "reserve collection." However, if the trait property is rare, perhaps even as rare as a unique allele for one single germplasm sample, then getting this allele (germplasm sample) included in the core selection subset is mostly a matter of luck. A completely random selection of accessions would here be expected to perform just as well as the core selection method. The purpose of the new trait mining method is to calculate a targeted subset to include accessions with a higher probability of holding the desired trait property. This FIGS subset rather than the core selection subset will then be used for the field trial experiment.

Focused Identification of Germplasm Strategy

This current study on Nordic barley landraces contributes to the further development of the trait mining methodology. Mackay and Street (2004) described FIGS as an alternative to the core selection strategy. The early origin of the FIGS theory dates back to 1995 when Michael Mackay was asked to provide wheat germplasm as a source to improve the boron tolerance of wheat crops in South Australia. Mackay knew that boron toxicity is more common in soil formed from marine sediments and selected some wheat landraces from former marine areas of the Mediterranean region in response to these seed requests. The selection of these wheat landraces proved successful and the train of thought leading to the development of the FIGS strategy was started (M. Mackay, personal communication, 2006). The overall goal of the new trait mining strategy is to improve utilization of genebank collections by increasing the chances of finding useful traits in relatively small subsets of germplasm (Mackay and Street, 2004).

Ecogeographic Association for Prediction of Missing Evaluation Data (1984)

Peeters and Williams (1984) published in the mid-1980s the suggestion to use ecogeographical principles to help in selecting genebank samples with missing evaluation data. Peeters et al. (1990) made a follow-up to this suggestion with an experiment to explore the association between salt tolerance and ecogeography, with the focus on rainfall. They found only a weak association between the ecological variables and the salt-tolerance responses.

Ecogeographic Studies with DIVA-GIS

The DIVA-GIS software (http://www.diva-gis.org) was introduced in 2001 as a free Geographic Information Systems (GIS) tool for documentation of genetic resources. DIVA-GIS includes useful tools to extract environmental variables for georeferenced genebank accessions and also features to analyze biodiversity data such as the prediction of ecogeographic distributions. The user manual for DIVA-GIS suggested that "traits can be related to ecological conditions at the places where the collections were made" (Hijmans et al., 2001, p. 17). Guarino et al. (2002) suggested that the association between environmental data for the germplasm location of origin (extracted by DIVA-GIS) and agronomic traits could be used in an ecogeographic gap analysis to guide complementary germplasm collecting expeditions to fill ecogeographic gaps in the genebank collections. Jarvis et al.

(2003) analyzed the association between ecogeographic data and the distribution of wild peanut (*Arachis* spp.). This was an important study to verify the applicability of the so-called ecological niche models (Elith and Graham, 2009) to predict the species distribution for these crop wild relatives.

Ecogeographic Predictability of Traits for Potato

Scientists at USDA and CIP have studied taxonomic, biogeographic, and ecogeographic predictability for traits in potato. The correlation between frost tolerance and the temperature of the origin locations for collected wild potato was modeled by Hijmans et al. (2003) using the GLM regression procedure with the SAS software package. The association of the trait expression (frost tolerance) with the ecogeographic factor (temperature) was identified, but concluded to be too weak to be useful. Jansky et al. (2006) reported weak and inconsistent predictive association between resistance to white mold [*Sclerotinia sclerotiorum* (Lib.) de Bary] and the ecogeography of the original collecting site for 144 accessions from 34 species of wild potato. They used a linear partial regression model including only the significantly correlated climate variables. The model explained on average less than 40% of the variation. Jansky et al. (2008) explored predictability of early blight in potato from the associations to taxonomic nomenclatural units and environmental factors. The most significant spatial and environmental variables were selected and analyzed as a linear regression tree. In this study the monthly average precipitation in July was identified as the most discriminating rule to predict resistance to early blight. In a recent study by Spooner et al. (2009), the association between ecogeographic variables and 32 different pest and disease ratings were calibrated with two predictive statistical methods, support vector machines (Guo et al., 2005), and random forest (Thuiller et al., 2003). Only 6 traits from the total of 32 pests and disease traits were successfully predicted from the environmental variables. The trait dataset they used was a compilation of reported resistance observations from many different sources and was combined with observations from multiple years ranging from 1986 to 2001. As Spooner et al. (2009) mention in the paper, the year-to-year variation is not taken into account, and this could be one explanation for the poor predictability from the computer models in this study.

Focused Identification of Germplasm Strategy Studies at the International Center for Agricultural Research in the Dry Areas

A recent study at the International Center for Agricultural Research in the Dry Areas (ICARDA) (Bouhssini et al., 2009) has demonstrated successful prediction of resistance to the Sunn insect pest (*Eurygaster integriceps* Puton) using the FIGS method. One new durum wheat and eight new bread wheat accessions with good levels of resistance to the Sunn pest were identified. A method of stratified multiple hierarchical principal component regression (PCR) clustering (with the SPSS software) was used to arrive at a FIGS set of 534 accessions where 9 novel sources of Sunn pest resistance were identified in field screening. Prior screening during 5 yr of a total of almost 2000 accessions from the same 16,000 wheat accessions did not result in the identification of any new sources of Sunn pest resistance (Street et al., 2008). A different study, including the same 16,000 wheat landraces, addressed the identification of resistance against the Russian wheat aphid (*Diuraphis noxia* Mordvilko). Using the FIGS method, a subset of 510 accessions was developed. Previously, several thousand wheat accessions had been screened at ICARDA for resistance against the Russian wheat aphid, with no success. From the FIGS subset, 12 accessions showed high to moderate resistance and 2 new resistance alleles were identified (Street et al., 2008). Another recent study performed at the University of Zurich, Switzerland applied the FIGS method for identification of novel alleles of resistance for powdery mildew in wheat (Bhullar et al., 2009). Starting from these 16,000 wheat accessions, a FIGS set of 1320 accessions was derived. Multivariate analysis methods were applied to identify collecting sites environmentally similar to the collecting sites for the accessions previously known to show good resistance to powdery mildew. Field screening of the FIGS subset of 1320 accessions resulted in the identification of 211 resistant accessions. Molecular analysis of these 211 accessions identified 7 previously unknown alleles for powdery mildew, doubling the number of known resistance alleles for the Pm3 locus (Bhullar et al., 2009). Recently an unpublished desktop study was performed at ICARDA to compare different subset selection methods from 5000 germplasm accessions screened for UG99 stem rust. Three subsets were derived by (i) the standard core selection strategy to maximize genetic diversity of the subset, (ii) as a set of 500 random selected accessions, and (iii) as a FIGS set of 500 accessions. The core set and the random set resulted in identification of roughly 10% resistant accessions, while the FIGS set resulted in 17.9% resistant accessions (K. Street et al., ICARDA, personal communication, 2009). These results demonstrate that the FIGS method is able to derive a manageable subset of material for screening with a higher probability for finding the desired trait than the normal core selection subset.

Multiway Data Structure

A multivariate dataset is often organized as a rectangular array of observations with two intersecting "modes" or "ways." The first mode is the records and the second mode is the variables. Many of these datasets have a three-way or higher order data structure. Often the same objects are observed for the same variables under different experimental conditions

(for example, locations and times). These conditions represent the third mode. Traditional factorial data analysis methods can only be applied to two-way data, and the multiway data structure is unfolded to a two-way matrix. Normally the first mode with the records is kept intact and the additional modes (with, for example, the conditions) are unfolded to the second mode with the variables. This will produce a so-called short and wide data array with many more columns than records. This reduction of the data structure often causes a significant loss of information. The two-way model will not include structured variation across the third mode (experimental conditions). Structured information between the entire so-called slabs or matrices of the first two modes (records and variables) repeated at the different instances of the third mode is lost (Bro, 1997).

Multiway Data Models

The multiway data structure is often studied with factorial analysis methods. In factorial analysis the structured information from correlated variables are condensed in so-called latent factors. These factors replace the actual variables in the data analysis and the number of factors is usually lower than the number of variables. With the multiway factorial methods used here, these factors are constructed so that they are othogonal to each other. The calibration of the model system can be seen as the rotation of the model system to find the best fit to the actual data. The calibration of the model is an iterative process toward the best fit. The Tucker model (Tucker, 1966) can be seen as a special case of the Principal Component Analysis (PCA) for multiway data structures. The PARAFAC model (Carroll and Chang, 1970; Harshman, 1970) is a special case of the Tucker model. The N-PLS regression method (Bro, 1996) is based in part on the PARAFAC method, and follows many of the same assumptions and characteristics as this model. Harshman and Berenbaum (1981) have summarized the particular assumptions of the PARAFAC model. (i) The factorial content of each slab (within each factor) is assumed to have proportional patterns of variation across the other ways of the data. For FIGS studies this assumption can be seen as a requirement that the ecogeographic variables at least roughly do the same thing at each of the landrace source locations. Next there is another assumption that is required for the valuable uniqueness of the PARAFAC solution (Kruskal, 1977), and thus the meaningful interpretation of the factors as explanatory factors (or latent real properties) (ii) Each factor should have a distinct (nonproportional) variation across the ways of the data.

Multiway Data Analysis

For this current study, a predictive computer model for the correlation between the trait dataset and the climate dataset was calibrated with the N-PLS regression method (Bro, 1996). The N-PLS method is a mixture of parallel factor analysis, PARAFAC (Carroll and Chang, 1970; Harshman, 1970) and partial least squares (PLS) (Wold, 1966). The N-PLS can be seen as a generalization of the PLS regression (PLS-R) (Wold et al., 2001) to multiway situations (Smilde, 1997). The multiway data analysis methods (like PARAFAC and N-PLS) have a number of benefits over the bilinear two-way matrices methods (like PLS, mutiple linear regression [MLR] and PCA). If the dataset is appropriate for a multiway structure (see assumptions above), then the modeling of the unfolded two-way matrices has a number of disadvantages. The bilinear (two-way) model consumes more degrees of freedom than a multilinear (multiway) model. The two-way models are more difficult to interpret, as they will display a mixture of modes/ways. With bilinear methods the multiway information is simply thrown away (as mentioned above). The two-way model will usually provide a closer fit to the calibration dataset, but will include more noise. The multiway methods will thus provide a more robust model with less overfit. Multiway methods most often show improved predictive power compared with two-way methods when the multiway data structure is appropriate (Bro, 1997). For more background information on multiway analysis, see the student textbook by Smilde et al. (2004).

MATERIALS AND METHODS

The input data for this study is a collection of three different datasets: (i) germplasm passport data, (ii) trait measurements, and (iii) environmental ecogeographic data.

(i) Germplasm Accessions (Passport Data)

The germplasm seed samples represent barley landraces (traditional cultivars) from the Nordic ex situ genebank collection (see Table 1). Longitude and latitude coordinates for the original source locations were extracted (or verified) from GIS software using a place name gazetteer. The gazetteers included place name searches with Google Maps (http://maps.google.com), BioGeomancer (Guralnick et al., 2006), Global Gazetteer Version 2.2 (http://www.fallingrain.com/world/; Falling Rain Genomics, 2010), the Getty Thesaurus of Geographic Names (http://www.getty.edu/research/conducting_research/vocabularies/tgn/), and the map section of the online phone book for some of the Nordic countries (http://map.krak.dk, http://kart.gulesider.no, and http://kartor.eniro.se; Eniro/Krak, 2010). All the landrace accessions are part of the Nordic Genetic Resources Center (NordGen) genebank collection and all seed samples are freely available to plant breeders or scientists worldwide on request. For more information on these accessions, see the SESTO Genebank Information and Management System (2009).

(ii) Trait Dataset (Characterization and Evaluation Data)

The barley trait dataset was developed by Kolodinska Brantestam (2005) and published as part of her doctoral thesis. The

Table 1. Landrace accessions included in this study.

Accession no.	Locality and country of origin	Elevation	Latitude	Longitude
		m†	——————decimal†——————	
NGB27	Sarkalahti, Luumäki, Finland	95	61.033	27.333
NGB456	Dønna, Nordland, Norway	71	66.117	12.500
NGB468	Trysil, Norway	400	61.283	12.283
NGB469	Bjørneby, Norway	400	61.283	12.283
NGB775	Överkalix, Allsån, Sweden†	45	66.400	22.933
NGB776	Överkalix, Sweden†	100	66.400	22.767
NGB792	Luusua, Kemijärvi, Finland	145	66.483	27.350
NGB2072	Finset, Norway	1220	60.600	7.500
NGB2565	Öland, Sweden	11	56.733	16.667
NGB4641	Støvring, Jylland, Denmark	55	56.883	9.833
NGB4701	Faroe Islands†	81	62.017	−6.767
NGB6300	Faroe Islands†	81	62.017	−6.767
NGB9529	Lyderupgaard, Denmark	9	56.567	9.350
NGB13458	Koskenkylä, Rovaniemi, Finland	91	66.517	25.867

†Two landraces are from very close locations in Överkalix, Sweden (NGB775, NGB776), and two landraces are from the Faroe Islands (NGB4701, NGB6300). These landrace origin sites could have issues of spatial autocorrelation. The other locations of landrace origin are well distributed and not likely to have spatial autocorrelation issues.

eld trials were performed in two replications during 2 yr (2002 and 2003), at three locations (Bjørke in southern Norway, Landskrona in southern Sweden, and Prieku i in Latvia). Six agronomic traits were scored: days to heading (from 1 June), days to maturity (from 1 July), plant height (cm), harvest index (percent), volumetric weight (kg/hl) and thousand-grain weight (gram). From the total trait dataset of 196 accessions, the 19 accessions classi ed as landraces were extracted for this trait mining analysis. From these 19 landraces, geo-referencing was successful for 14 accessions. For each of these 14 accessions there are a total of 12 trait measurements (3 locations, 2 yr, 2 replications) for each of the 6 traits (a grand total of 1008 trait observations). The trait dataset is included in Appendix 1.

(iii) Environmental Data

Ecogeographic data were extracted for the source locations where the accessions (landraces) were originally developed and collected. The environmental variables include the monthly mean values for daily minimum and maximum temperatures, and the monthly mean for daily precipitation. Climate data was extracted from the WorldClim dataset (Hijmans et al., 2005) using the DIVA-GIS software (Hijmans et al., 2004). Data from the spatial grid with the resolution of 30 arc seconds (approximately 1-km grid) was used for this study. The WorldClim environmental dataset was downloaded from http://www.worldclim.org/.

Data Analysis

From the initial explorative principal component analysis (PCA), one observation was identi ed as a strong outlier. This observation was for landrace NGB6300 scored for harvest index (Trait 4) at Prieku i (Latvia) in 2003 for replicate 2. This score value showed a very high leverage in the plots of Hotelling T-square against residuals from the initial PCA analysis. For the data analysis reported here, this single data point was removed (set as missing value). The models produced with this landrace (NGB6300) excluded all together showed similar results to the models where only the outlier data point was excluded.

Multiway Data Design

The environmental dataset and the crop trait dataset were each organized in a three-way array (also called a data cube) as illustrated by Fig. 1. Each of the two three-way arrays was analyzed with the PARAFAC method. The N-PLS algorithm was in the next step applied to calibrate a regression model between the ecogeographic dataset organized as a multiway structure (as the independent variables), and the trait data organized as a set of one-way structures (vector) as the dependent variable. Different N-PLS models were explored including tri-PLS2 (multiple dependent variables, crop traits) and tri-PLS1 (one single dependent variable, crop trait). The tri-PLS2 models were more di cult to interpret, it was also more di cult to decide the number of optimal components (factors), and they did not perform better than the tri-PLS1 models. Only the results from the tri-PLS1 method (N-PLS) are reported below. The trait dataset includes trait data measured under 6 di erent experimental conditions for 6 di erent trait characters. For the regression data analysis with the tri-PLS1 method below, the complete dataset was split into 36 subsets (6 traits, 3 experiment locations, and 2 experiment years).

Preprocessing (Centering and Scaling)

The environmental dataset and trait datasets were preprocessed with the autoscaling method (Varmuza and Filzmoser, 2009) before the data analysis. The autoscaling preprocessing method is a combination of mean centering and variance scaling. Mean centering means that for each dataset the mean for each variable is subtracted from the individual data values. Mean centering removes the absolute intensity information. This preprocessing strategy is applied to avoid the model focusing on the variables with the highest numerical values (intensity). Variance scaling means that each data value is divided by the standard deviation of the data series (record). After autoscaling, all variables will have a mean of zero and a standard deviation of one. The PLS Toolbox software (npreprocess routine) was used for appropriate multiway preprocessing (Bro and Smilde, 2003; PLS-Toolbox, 2009).

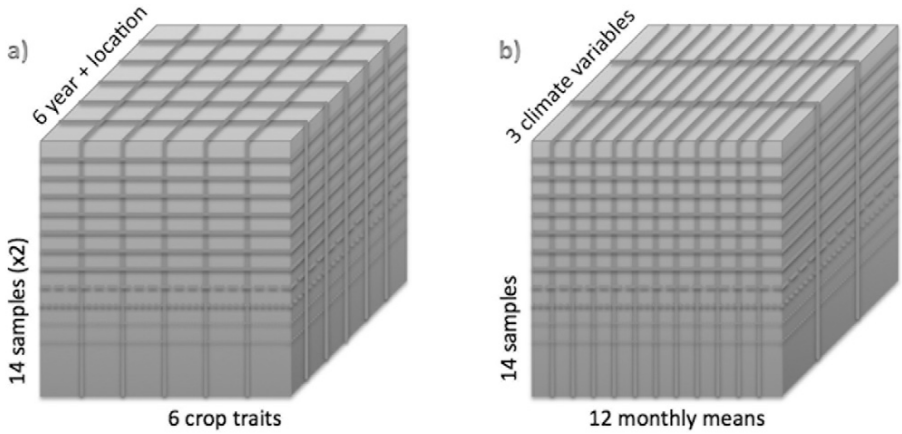


Figure 1. Multiway data design. The crop trait data (a) was organized as tri-linear (three-way) cube with 28 landrace measurements (14 samples, 2 replications) by 6 crop trait properties by 6 experimental conditions (2 yr, 3 locations). The environmental data (b) was organized as a trilinear (three-way) cube with 14 landrace origin locations by 12 monthly means by 3 climate variables.

Parallel Factor Analysis (Split-Half Analysis)

A PARAFAC split-half analysis (Harshman and Lundy, 1984) was made to find the appropriate multiway data design and to assess the stability of the corresponding multilinear models. Because the PARAFAC model produces a unique solution (Kruskal, 1977), the model parameters for a stable situation will have the same profile for the two split-half sets, and for the model for the complete set. Both the trait dataset and the climate dataset were structured as a three-way cube for the respective split-half analysis. The dataset was in the first step divided by random into two halves. One PARAFAC model was calibrated for the complete dataset and one model for each of the two halves. Finally the model parameters for these three models were compared.

Predictive Performance from Multilinear Partial Least Squares, Multilinear Regression

The trait scores predicted from the N-PLS model were compared with the actual measured trait scores, and the performance of the predictive model evaluated using the Pearson product-moment correlation coefficient (r) and reported as coefficient of determination (r^2). The predictions were from the cross-validation (leave-one-out) method. Each landrace was, of course, predicted from a new N-PLS model calibrated without this sample included. And both of the two replicate measurements for the landrace samples were, of course, taken out together. The total number of 14 samples (landraces) was considered too low to create a separate independent test set.

Software

Multivariate analyses were performed using MATLAB (MATLAB, 2009) and the PLS Toolbox for MATLAB (PLS-Toolbox, 2009). MATLAB is available online from http://www.mathworks.com/, and the PLS Toolbox is available online from http://www.eigenvector.com/.

RESULTS

Parallel Factor Analysis (Split-Half Analysis)

For the trait data, different split-half simulations resulted each time in the same solutions. For the climate dataset, the split-half results were somewhat less stable (produced the same unique solution in 2 out of 10 runs). However, both datasets produce evidence for the acceptable stability in the multiway data structures for PARAFAC decomposition. Even with as few as the 14 landraces included in this study, the multiway methods (including N-PLS) are thus still expected to calibrate stable solutions.

Multilinear Regression

All of the predictive computer models (N-PLS) showed acceptable normal distributions for the residuals (root mean square error from cross-validation). The so-called influence plot of the Hotelling's T-square (Hotelling, 1931) against the residual for each landrace did not show any of the landrace samples as isolated from the other samples. Landrace NGB456 originating from Dønna in Northern Norway showed typical high leverage and large residuals, but without isolation from the other samples (and the computer models with NGB456 removed also showed slightly lower predictive performance). The estimated variance explained by the model was typically much higher for the independent variables, environmental data (average 93%), than for the dependent variables, trait data (average 67%). The summary of results from the computer models (N-PLS) is reported in Table 2. The strength of the association between the environment and the crop traits is reported as the coefficient of determination (r^2) between the predicted trait score and the observed trait score.

Results for Trait 6, Grain Weight

The predictions for Trait 6 (grain weight) show throughout no correlation to the measured values whatsoever. For four out of the six experiment subsets, the predictions for this

Table 2. Summary of predictive performance, coefficient of determination (r^2_{cv}) from multilinear partial least squares (N-PLS).

Experiment location and year	Trait 1 heading days	Trait 2 ripening days	Trait 3 length of plant	Trait 4 harvest index	Trait 5 volumetric weight	Trait 6 grain weight†
Prieku i 2002‡	0.16	0.10	0.11	0.16	0.70***	–
Prieku i 2003	0.65***	0.14	0.55**	0.40**	0.66***	0.00
Bjørke 2002	0.00	0.32*	0.38**	0.59***	0.51**	0.01
Bjørke 2003	0.42**	0.65***	0.66***	0.31*	0.24*	–
Landskrona 2002	0.40**	0.68***	0.04	0.55**	0.26*	–
Landskrona 2003§	0.01	0.43**	0.01	0.01	0.33**	–

*, **, ***Significant at the 0.05, 0.01, 0.001 probability levels, respectively. The critical values for the coefficient of determination are calculated for one-tailed test (positive correlation only) of significance for the model predictions of the 14 landraces (12 degrees of freedom). The coefficients for the negative correlations are not reported, but designated as a dash (–).

†Trait 6 (grain weight) was not predicted by the N-PLS regressions.

‡The climatic conditions for the field trials at Prieku i in 2002 was considered too extreme and caused abnormal growth development for many of the plants.

§The climatic conditions for the field trials at Landskrona in 2003 was considered too extreme and caused abnormal growth development for many of the plants.

trait were even slightly negatively correlated to the true values (negative correlation coe cient, r). The coe cient of determination (r^2) for positive correlations for this trait was never higher than 0.01 (Bjørke in 2002). The association to the environmental data for this trait is very di erent from the other traits (Table 2, †), and was removed from the nal results as will be discussed further below.

Results from Prieku i, 2002 and Landskrona, 2003

The predictive computer models did not explain the trait scores from Prieku i in 2002 and Landskrona in 2003. The local weather conditions during these two eld seasons seem to be too extreme for normal trait expression. Even the normal growth development was observed to be disturbed for many of the plants. For Prieku i 2002, May was unusually dry followed by an unusually wet June as can be seen from Table 3 († and ‡). The wet June caused germination on the spikes in the eld for many of the earlier varieties. The predictions for the trait scores measured at Prieku i in 2002 have coe cients of determination (r^2) between 0.10 and 0.16, except for Trait 5 (volumetric weight) with r^2 = 0.70 (Table 2, ‡). For Landskrona 2003, June was unusually dry as seen from Table 3 (§). This dry period was during the critical growth stage of grain lling. The coe cient of determination for this experiment site was close to zero (r^2 = 0.01), except for the traits ripening days at r^2 = 0.43 and volumetric weight at r^2 = 0.33 (Table 2, §). Results from Prieku i in 2002 and Landskrona in 2003 were removed from the nal results as will be discussed further below.

DISCUSSION

Predictive associations were found between ecogeography and morphological crop traits for Nordic barley landraces for ve of the six traits analyzed. The performance of the predictive computer models varied signi cantly between the different morphological traits. The predictive performance also proved to be sensitive to the local environmental weather conditions during the eld experiments. For two of the eld experiments (at Prieku i in 2002 and at Landskrona in 2003) the environmental conditions during the growth season were so extreme that the plants showed abnormal growth development (Kolodinska Brantestam, 2005). The heading days and the ripening days were measured during the growth season, the other traits after harvest. At Prieku i in 2002, the heading occurred between 22 June and 9 July. At Landskrona in 2003, the heading occurred between 13 June and 29 June. The predictive models for the trait heading days showed a lower predictive performance than most of the other traits. It is possible that this trait, measured earlier in the growth season, is more sensitive to the in uence from the local weather conditions during the trial season.

Grain Weight, Seed Size

The trait for grain weight (Trait 6) showed absolutely no association with the ecogeography at the landrace location of

Table 3. Local climate at the experiment site during the growth season, precipitation.

Experiment location and year	Sowing week	May	June	July	August
		mm			
Prieku i 2002	17	38.2†	111.1‡	67.0	11.3
Prieku i 2003	19	88.0	59.2	87.8	175.8
Bjørke 2002	17	82.9	67.4	128.5	136.5
Bjørke 2003	21	75.1	85.7	67.1	53.2
Landskrona 2002	13	53.5	75.3	76.4	68.9
Landskrona 2003	15	70.7	40.4§	76.0	45.7

†The growth conditions at Prieku i in 2002 showed an unusually dry May. This caused many of the plants to germinate in the spikes.

‡The growth conditions at Prieku i in 2002 showed an unusually wet June. This caused many of the plants to germinate in the spikes.

§The growth season at Landskrona in 2003 showed an unusually dry June, causing failed grain filling for many plants.

origin. Landraces are developed by traditional farming at the location of origin and express traits adapted to the local ecogeography from this area. The source germplasm selected to become a landrace is likely to have some traits already suited to the local ecogeography of the area. Some of the traits for a landrace are also developed and shaped further by the local ecogeography. Other traits are more dominantly the result of adaption to human selection pressure to suit the farmer rather than adaption to t the environmental conditions (Darwin, 1859). There are also traits that are selected by the conditions created by the agricultural management practices rather than as the result of an active conscious selection from the farmer (Harlan et al., 1973). It is possible that the grain weight is a trait with a more dominant human or agronomic selection and thus less associated with the local ecogeography. Harlan et al. (1973) list increased seed size as one of the characteristic domestication traits. Fuller (2007) describes the increase in seed size as one of the key archaeological indicators of early domestication in cereals. Giles (1990) describes the paradox of measured increase in tness from larger seeds, yet no sign of natural selection toward reduced variation in seed size (and larger seeds) in wild barley. A recent study by Gambin and Borras (2010) discusses these paradoxes on the evolutionary adaption of seed size further. The mechanism behind selection and inheritance of seed size remain at least to some degree not yet fully understood. The results from this current study suggest that the grain weight has only a very weak (if any) association with the ecogeography of the location of origin in barley landraces.

Different Experimental Conditions

A common formula often used to explain the measured trait expression is that the phenotype is determined by the genotype e ect plus the environment e ect plus the e ect of the genotype and environment interaction (G×E). The G×E e ect is known to cause di erent genotypes to sometimes rank di erently between the test environments (so-called crossover e ects) (Baker, 1988). In many studies analyzing crop trait datasets, the unreplicated data points for trait measurements across di erent experiment locations and years are simply combined and analyzed together. It is, of course, valuable to include observations of the same germplasm accessions for trait performance across di erent environments (experiment locations and years). Such multilocation trials are particularly important for the identi cation of the so-called crossover G×E e ects (Baker, 1988). However, sometimes composite genebank trait datasets are developed with observations from di erent locations and sometimes as few as one single observation of each germplasm. It is worth considering that under very di erent experimental conditions, the experimental observations could measure entirely di erent qualities (or traits) of the germplasm accession. The G×E e ect could thus change the trait into another trait with a di erent genetic background, even if the experimental trait protocol is the same. The results from this current study suggest that care should be taken when the trait dataset includes measurements from di erent experimental conditions. In this current experiment, the conditions during two of the subexperiments (at Prieku i in 2002 and at Landskrona in 2003) were so extreme that this a ected the normal plant growth. In these two situations it seems that the experimental conditions were so extreme that something di erent from the desired trait was measured. The estimated predictive model performance also shows signi cant variation between di erent subexperiments (eld location and year). This variation in predictive performance could perhaps be a symptom of the G×E e ect.

With appropriate data analysis methods and appropriate dataset design, the relevant e ect from the di erent conditions during di erent subexperiments could perhaps be isolated. If data points (records) from di erent subexperiments are simply appended together to the same dimension of the dataset, the e ects of the di erent experimental conditions will be mixed and not included in the data analysis models. This would introduce variation that often could have been explained better if the multiway nature of the trait dataset were preserved. When building the reported computer models for this current study, only the multiway design of the ecogeographic environmental dataset was maintained. The trait dataset was split into new subsets based on the experimental conditions. Alternative computer models for other data array designs were explored, but not reported here. In one such alternative experiment, the climate data was organized as a three-way data cube, the very same design as for the reported models, and the trait data was here organized as a two-way matrix. The two-way trait data matrix included landrace samples as records and the di erent traits as columns. The three-way environmental dataset and the two-way trait dataset were tted to a tri-PLS2 regression model. Systematic variation across di erent traits would here be preserved and isolated as a separate dimension. The predictive performance for these tri-PLS2 models was however more or less the same as for the tri-PLS1 models, but the visualization and the interpretability were dramatically reduced. It is possible to imagine predictive modeling for the associations between even higher-order data cubes. The trait data cube could be structured with independent dimensions for the di erent experiment locations and years. These dimensions for the subexperiments could also include attributes for the local environmental weather conditions. We would expect to nd an underlying systematic variation across such subexperiments (G×E), and that this variation could contribute to explaining the association between the trait expression and the ecogeography of the landrace origin. Such higher order N-PLS models could perhaps provide a good method to model the G×E e ect. The development of higher order N-PLS methods and software implementations are still under development.

Anticipated future advances here could prove particularly valuable for ne-tuning and extracting even more information from the trait mining methods.

Multiway Data Analysis Method

This current study explores the suitability of novel multilinear multivariate analysis methods for the prediction of economically valuable crop trait properties from ecogeographic environmental data from the location of crop origin. The most important aspect of this method compared with other methods used for ecological niche models, is that the ecogeographic variables are maintained in a multiway structure instead of unfolding to a bilinear matrix. Two-way analysis methods often require unrealistic assumptions, such as statistical independence between the environmental variables. The problems of covariance between these variables will be signi cantly reduced by the application of multiway analysis methods (Bro, 1996). The multiway PARAFAC and N-PLS models are more robust to noise, and usually provide improved interpretation over two-way methods (Gurden et al., 2001). The multiway methods most often produce a lower t of the models to the data, but also a much lower over t. In other words, the multiway methods focus on the information and leave more noise out of the model. The speci c multiway analysis with the PARAFAC and the N-PLS methods o er yet another signi cant advantage compared with other multilinear methods in that the calibrated solution is unique. There will theoretically be only one unique N-PLS solution for the calibration of the regression model for a given dataset. The uniqueness of the solution improves the interpretability and visualizations of the model results. The basic and practical problem can however often be to nd the appropriate multilinear design for the dataset (data cube design). For this current study a four-way multilinear structure for the climate dataset was rst explored, but did not calibrate to a stable PARAFAC model. The advantages of the multiway methods are even more valuable for datasets with few samples. The unfolding of the multiway data to a bilinear matrix give you a so-called short and wide matrix, and only as many examples for the calibration step of the regression as the available number of samples. Keeping the multiway data design for the multilinear N-PLS regression, the calibration step will have signi cantly more examples to work with (Bro, 1996). In this study the number of samples (landraces) are on the lower end, and the advantage of using a multilinear regression method is thus particularly valuable. The positive results from the PARAFAC split-half analysis verify that the PARAFAC model seems to nd enough systematic variation in these few samples to calibrate a relevant model. The N-PLS regression model is based on the PARAFAC calibration routine and is thus also expected to calibrate a relevant model.

Computer Simulation Model

The current study presented here is a computer simulation study. A subset of the dataset was hidden from the data model as a validation test set, and the performance of the data model assessed by comparing the predictions of the hidden trait scores to the real values. In a real-life scenario the FIGS predictions will produce a FIGS subset to be screened for the target trait property in a eld trial. To compare the FIGS predictions to the actual values, the eld trial experiment needs to be completed. The results obtained from a computer simulation study are likely to give more optimistic predictions than a real-life FIGS experiment because the validation test set is more similar to the training set used to calibrate the FIGS model. When performing a real-life FIGS study, special attention needs to be given to the potential extrapolation problems. The trait dataset and the germplasm samples used for training and calibration of the computer model must be representative for the larger dataset of germplasm samples the FIGS model will be applied to. This current study was performed with a very low number of germplasm samples. The full potential of the described FIGS approach with multiway computer-intensive methods is not illustrated with this small dataset. If the genebank accessions are well georeferenced, then feeding the FIGS model with a larger dataset would lead to more computer time, but not signi cantly more human work input. This FIGS approach would thus be expected to scale well to the problem of nding a targeted small subset in a large genebank collection.

Ecogeographic Variables

The trait mining analysis as described here is expected to improve with the careful selection of the most relevant independent variables (ecogeography). Expert knowledge on the most predictive ecogeographic environmental variables can also be used to improve the model predictive performance. However, there is today limited access to high quality global climate variables, or at least to climate data that cover the entire geographic area of a study. Current rapid advances in physical geography provide new relevant environmental variables that should be considered in future trait mining models. Precision of the ecological niche and the trait mining methods are expected to further improve with the advance of available ecogeographic climate variables.

Similarity to Species Distribution Models

The FIGS method can be seen as building an ecological niche model for the expression of the target trait property. The analysis of the association between the source location for traditional cultivars and trait experiments for these landraces can be compared to nding a so-called fundamental ecological niche (Hutchinson, 1957) where genetic resources with the target trait property are likely

to be found. Some of the early trait mining studies follow a method similar to the climate envelope models as implemented by the early environmental niche modeling algorithms for the prediction of species distribution such as BIOCLIM (Busby, 1991) and DOMAIN (Carpenter et al., 1993). In several studies comparing the performance of different ecological niche modeling methods, the early envelope models were outperformed by the more complex novel methods (Elith et al., 2006; Guisan et al., 2007; Elith and Graham, 2009). Most ecological niche modeling methods take as input the observed presence and absence of the species at georeferenced locations. The expected binary data input for most ecological niche modeling methods would pose a limitation for utilization in trait mining. The multiway data analysis methods applied in this current study have not yet been explored for the prediction of species distributions. Jarvis et al. (2003) provides an early example of the species distribution models applied to plant genetic resources.

Allele Mining (Other Uses)

Modern molecular genetics methods are also more often applied to detect agronomically valuable alleles in seed bank collections (Tanksley and McCouch, 1997). See, for example, Prada (2009) for a recent description of how the new allele mining strategies are developed to search landraces and crop wild relatives for valuable allelic variants lost during domestication. However, these allele mining methods have some limitations when searching for germplasm with a specific agronomic trait valuable for crop improvement. Allele mining requires understanding of the allele(s) involved in the expression of a desired trait. Available experimental molecular analysis data for the candidate germplasm is also required. The trait mining strategy as described above could perhaps provide a simulated prescreening to select germplasm for a more in-depth allele mining experiment. Hübner et al. (2009) describes a method where the ecogeographic data (temperature and precipitation) is successfully associated to the population structure in wild barley from 51 collection sites in Israel, as estimated with a total of 42 simple sequence repeat (SSR) markers. Similar future studies of associations between data from molecular genetics and ecogeographic parameters could guide the understanding of allele functions.

Landraces—Adaption to Origin Location

Landraces are developed during long-term traditional cultivation at the same location. Each landrace is exposed to a human selection pressure as well as the selection pressure from the environmental conditions (ecogeography) of the location where the traditional cultivation takes place. Some evidence also suggests an adaptive evolution of primitive crops and landraces (Gepts and Papa, 2002). The predictive patterns in the climate variables could, however, also very well be there from reasons other than adaption to the environment in the individual landraces. Traditional farming will keep the crop material best suited for the local conditions. During the cultivation of the material, the farmer will also actively select crop material for improved adaptability to these local conditions. The trait mining strategy described here does not require the association between the ecogeography and the trait expression to be from adaptive evolution of primitive crops and landraces. The trait mining method does not discriminate predictive association caused by adaption from predictive association caused by the selection made by the farmer of suitable original source material for the founding of the landrace.

Limitations to the Use of Plant Genetic Resources

Marshall (1989) acknowledges the lack of documentation and description of the material as an important limitation to the use of plant genetic resources (PGR). Marshall (1989) does however warn against meeting this demand with a systematic evaluation of collections, and argue the lack of prebreeding preceded by more selective, targeted evaluation as the most important limitation to the use of gene bank collections. Esquinas-Alcázar (2005) discusses the lack of information on genebank accessions remaining still today as one of the most important limiting factors to utilization of plant genetic resources. Chapter 3.8 and 4.7 of the recent Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture (FAO, 2009) points to the lack of relevant documentation for the world's ex situ plant genetic resources as one of the most important limitations to the improved use in crop improvement and research. Chapter 4.3 from this report also points to the almost unanimous suggestion from the reporting countries that lack of adequate characterization and evaluation data is the most significant obstacle to greater use of plant genetic resources for food and agriculture (PGRFA). Many countries also mention increasing use of core collections and other collection subsets as an efficient strategy to produce the demanded trait data. The lack of adequate information on accessions is however reported as the major obstacle to the establishment of core collections (FAO, 2009).

CONCLUSIONS

The proposed new trait mining method or focused identification of germplasm strategy (FIGS) assumes predictive association between the ecogeography of the germplasm origin location and the trait scores. This association is expected to be stronger in more primitive cultivars than in modern high-yielding cultivars. Landraces are to a much larger part selected or adapted to fit the environment of their location of origin. Modern cultivars are developed for high yield and seldom targeted at a particular location, as is the case for the landraces. Different traits can, as

suggested by this current study, show very different association to the environment of the location of origin. Trait mining is here proposed as an alternative method to the core selection method. This current study suggests that the new trait mining method can provide an efficient and powerful alternative to the core selection method. It is, however, useful to remember the assumptions and limitations of the trait mining method. When the prerequisites for the trait mining method are not fulfilled, the core selection method is still the most appropriate method.

Supplemental Information Available

Appendix 1 is available online at http://www.crops.org/publications/cs.

Acknowledgments

Drs. Agnese Kolodinska Brantestam very kindly provided the original data from her study (Kolodinska Brantestam, 2005). Agnese Kolodinska Brantestam also contributed weather data for each of the experiment sites. Thanks also to other colleagues at the Nordic Genetic Resources Center (NordGen) for feedback and support during the research for this manuscript. Thanks to associate professor Dvora-Laiô Wulfsohn and professor Brian Grout from Copenhagen University, Faculty of Life Sciences, for valuable advice and feedback on the draft manuscript. Thanks to professor Rasmus Bro for teaching me multilinear data analysis and how to apply the multiway methods. Rasmus Bro is the originator of the multilinear regression methods applied in this study, and also shared with me his MATLAB scripts (Bro, unpublished results, 2009) to efficiently perform the split-half PARAFAC evaluation. Thanks to Michael Mackay (Bioversity) and Ken Street (ICARDA) for advice and background information for the FIGS method. This research was funded by a grant from the Nordic Genetic Resources Center (NordGen) and also supported by Bioversity International. Many thanks to the late Dr. Bent Skovmand for valuable help and advice when I was starting this research project.

References

Baker, R.J. 1988. Tests for crossover genotype-environmental interactions. Can. J. Plant Sci. 68:405–410.

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. Proc. Natl. Acad. Sci. USA 106(23):9519–9524 doi:10.1073/pnas.0904152106.

Bouhssini, M., K. Street, A. Joubi, Z. Ibrahim, and F. Rihawi. 2009. Sources of wheat resistance to Sunn pest, Eurygaster integriceps Puton, in Syria. Genet. Resour. Crop Evol. 56(8):1065–1069.

Bro, R. 1996. Multiway calibration. Multilinear PLS. J. Chemom. 10:47–61.

Bro, R. 1997. PARAFAC: Tutorial & applications. Chemom. Intell. Lab. Syst. 38:149–171.

Bro, R., and A.K. Smilde. 2003. Centering and scaling in component analysis. J. Chemom. 17:16–33 doi:10.1002/cem.773.

Brown, A.D.H. 1989. The case for core collections. In A.H.D. Brown et al. (ed.) The use of plant genetic resources. Cambridge Univ. Press, Cambridge, UK.

Busby, J.R. 1991. BIOCLIM–a bioclimate prediction system. p. 4–68. In C.R. Margules and M.P. Austin (ed.) Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Melbourne, Australia.

Carpenter, G., A.N. Gillison, and J. Winter. 1993. DOMAIN: A flexible modeling procedure for mapping potential distributions of plants and animals. Biodivers. Conserv. 2:667–680.

Carroll, J.D., and J. Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Eckart-Young' decomposition. Psychometrika 35:283–319 doi:10.1007/BF02310791.

Darwin, C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life (1st edition). John Murray, London.

Elith, J., C.H. Graham, R.P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R.J. Hijmans, F. Huettmann, J.R. Leathwick, A. Lehmann, J. Li, L.G. Lohmann, B.A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J.McC.M. Overton, A.T. Peterson, S.J. Phillips, K. Richardson, R. Scachetti-Pereira, R.E. Schapire, J. Soberón, S. Williams, M.S. Wisz, and N.E. Zimmermann. 2006. Novel methods improve prediction of species' distributions from occurrence data. Ecography 29(2):129–151 doi:10.1111/j.2006.0906–7590.04596.x.

Elith, J., and C.H. Graham. 2009. Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models. Ecography 32(1):66–77 doi:10.1111/j.1600–0587.2008.05505.x.

Eniro/Krak. 2010. Map sections from some of the online Nordic phone books available at http://map.krak.dk (Denmark), http://kartat.eniro. (Finland), http://kart.gulesider.no (Norway), http://kartor.eniro.se (Sweden) (verified 1 Aug. 2010). Eniro/Krak. Stockholm, Sweden.

Esquinas-Alcázar, J. 2005. Science and society: Protecting crop genetic diversity for food security: Political, ethical and technical challenges. Nat. Rev. Genet. 6:946–953 doi:10.1038/nrg1729.

FAO. 2009. Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Final draft version (CGRFA-12/09/Inf.7 Rev.1). Available at ftp://ftp.fao.org/docrep/fao/meeting/017/ak528e.pdf (verified 27 July 2010). Food and Agriculture Organization of the United Nations, Rome.

Frankel, O.H. 1984. Genetic perspectives of germplasm conservation. p. 161–170. In W. Arber, K. Llimensee, W.J. Peacock, and P. Starlinger (ed.) Genetic Manipulation: Impact on Man and Society. Cambridge Univ. Press, Cambridge, UK.

Frankel, O.H. 1989. Principles and strategies of evaluation. p. 105–120. In A.D.H. Brown, O.H. Frankel, D.R. Marshall, and T.J. Williams (ed.) The use of plant genetic resources. Cambridge Univ. Press, Cambridge, UK.

Frankel, O.H., and E. Bennett. 1970. Genetic Resources in Plants–their Exploration and Conservation. IBP Handbook No. 11. Blackwell Scientific Publications, Oxford, UK.

Frankel, O.H., and A.H.D. Brown. 1984. Current plant genetic resources: A critical appraisal. p. 1–11. In V.L. Chopra, B.C. Joshi, R.P. Sharma, and H.C. Bansal (ed.) Genetics: New Frontiers Vol. IV. Proc. of the Int. Congr. of Genet., 15th, New Delhi, India. 12–21 Dec. 1983. Oxford and IBH Publishing Co., New Delhi.

Fuller, D.Q. 2007. Contrasting patterns in crop domestication and domestication rates: Recent archaeobotanical insights from

the old world. Ann. Bot. (Lond.) 100:903–924 doi:10.1093/aob/mcm048.

Gambin, B.L., and L. Borras. 2010. Resource distribution and the trade-o between seed number and seed weight: A comparison across crop species. Ann. Appl. Biol. 156:91–102 doi:10.1111/j.1744–7348.2009.00367.x.

Gepts, P. 2006. Plant genetic resources conservation and utilization: The accomplishments and future of a societal insurance policy. Crop Sci. 46:2278–2292 doi:10.2135/cropsci2006.03.0169gas.

Gepts, P., and R. Papa. 2002. Evolution during domestication. In Encyclopedia of Life Sciences. Nature Publishing Group, London.

J. Paul Getty Trust. 2010. Getty Thesaurus of Geographic Names. Available at http://www.getty.edu/research/conducting_research/vocabularies/tgn/ (veri ed 1 Aug. 2010). J. Paul Getty Trust. Los Angeles, CA.

Giles, B.E. 1990. The e ects of variation in seed size on growth and reproduction in the wild barley Hordeum vulgare ssp. spontaneum. Heredity 64:239–250.

Falling Rain Genomics. 2010. Global Gazetteer version 2.2, Directory of Cities and Towns in the World. Available at http://www.fallingrain.com/world/ (veri ed 1 Aug. 2010). Falling Rain Genomics, Inc. Palo Alto, CA.

Google, Inc. 2010. Google Maps. Available at http://maps.google.com/ (veri ed 1 Aug. 2010). Google Inc., TerraMetrics Inc. and Tele Atlas BV. Mountain View, CA.

Guarino, L., A. Jarvis, R.J. Hijmans, and N. Maxted. 2002. Geographic Information Systems (GIS) and the Conservation and Use of Plant Genetic Resources. p. 387–404. In J.M.M. Engels, V.R. Rao, A.H.D. Brown, and M.T. Jackson (ed.) Managing Plant Genetic Resources. IPGRI, Rome.

Guisan, A., N.E. Zimmermann, J. Elith, C.H. Graham, S. Phillips, and A.T. Peterson. 2007. What Matters for Predicting the Occurrences of Trees: Techniques, Data, or Species' Characteristics? Ecol. Monogr. 77(4):615–630 doi:10.1890/06–1060.1.

Guo, Q., M. Kelly, and C.H. Graham. 2005. Support vector machines for predicting distribution of sudden oak death in California. Ecol. Modell. 182:75–90 doi:10.1016/j.ecolmodel.2004.07.012.

Guralnick, R.P., J. Wieczorek, R. Beaman, R.J. Hijmans, and the BioGeomancer Working Group. 2006. BioGeomancer: Automated Georeferencing to Map the World's Biodiversity Data. PLoS Biol. 4(11): E381. doi:10.1371/journal.pbio.0040381 BioGeomancer available at http://biogeomancer.org (veri ed 27 July 2010).

Gurden, S.P., J.A. Westerhuis, R. Bro, and A.K. Smilde. 2001. A comparison of multiway regression and scaling methods. Chemom. Intell. Lab. Syst. 59:121–136.

Harlan, J.R., J.M.J. DeWet, and E.G. Price. 1973. Comparative evolution of cereals. Evolution 27:311–325.

Harshman, R.A. 1970. Foundations of the PARAFAC procedure: Models and conditions for an "explanatory" multi-modal factor analysis. UCLA Working Papers in Phonetics 16:1–84.

Harshman, R.A., and S.A. Berenbaum. 1981. Basic concepts underlying the PARAFAC- CANDECOMP three-way factor analysis model and its application to longitudinal data. p. 435–459. In D.H. Eichorn, J.A. Clausen, N. Haan, M.P. Honzik, and P.H. Mussen (ed.) Present and past in middle life. Academic Press, New York.

Harshman, R.A., and M.E. Lundy. 1984. The PARAFAC model for three-way factor analysis and multidimentional scaling. In H.G. Law, C.W. Snyder, J.A. Hattie, and R.P. McDonald (ed.) Research methods for Multimode data analysis. Praeger, New York.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis. 2005. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climatol. 25:1965–1978. WorldClim climate data downloaded from http://www.worldclim.org/ (accessed 9 Jan. 2009; veri ed 1 Aug. 2010).

Hijmans, R.J., L. Guarino, C. Bussink, P. Mathur, M. Cruz, I. Barrentes, and E. Rojas. 2004. DIVA-GIS. Version 5.0. A geographic information system for the analysis of species distribution data. Manual available at http://www.diva-gis.org (veri ed 1 Aug. 2010).

Hijmans, R.J., L. Guarino, M. Cruz, and E. Rojas. 2001. Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. Plant Genet. Resour. Newsl. 127:15–19.

Hijmans, R.J., M. Jacobs, J.B. Bamberg, and D.M. Spooner. 2003. Frost tolerance in wild potato species: Unraveling the predictivity of taxonomic, geographic, and ecological factors. Euphytica 130:47–59.

Hotelling, H. 1931. The generalization of Student's ratio. Ann. Math. Stat. 2:360–378 doi:10.1214/aoms/1177732979.

Hutchinson, G.E. 1957. Concluding remarks. Cold Spring Harbor Symp. Quant. Biol. 22:415–427 doi:10.1101/SQB.1957.022.01.039.

Hübner, S., M. Hö ken, E. Oren, G. Haseneyer, N. Stein, A. Graner, K. Schmid, and E. Fridman. 2009. Strong correlation of wild barley (Hordeum spontaneum) population and precipitation variation. Mol. Ecol. 18:1523–1536 doi:10.1111/j.1365–294X.2009.04106.x.

Jansky, S.H., R. Simon, and D.M. Spooner. 2006. A test of taxonomic predictivity: Resistance to white mold in wild relatives of cultivated potato. Crop Sci. 46:2561–2570.

Jansky, S.H., R. Simon, and D.M. Spooner. 2008. A test of taxonomic predictivity: Resistance to early blight in wild relatives of cultivated potato. Phytopathology 98:680–687.

Jarvis, A., M.E. Ferguson, D.E. Williams, L. Guarino, P.G. Jones, H.T. Stalker, J.F.M. Vallis, R.N. Pittman, C.E. Simpson, and P. Bramel. 2003. Biogeography of wild Arachis: Assessing conservation status and setting future priorities. Crop Sci. 43:1100–1108.

Kolodinska Brantestam, A. 2005. A century of breeding—is genetic erosion a reality? Doctoral diss. Dep. of Crop Science, SLU. Acta Universitatis Agriculturae Sueciae Vol. 2005:30.

Kolodinska Brantestam, A., R. von Bothmer, C. Dayteg, I. Rashal, S. Tuvesson, and J. Weibull. 2004. Inter simple sequence repeat analysis of genetic diversity and relationships in cultivated barley of Nordic and Baltic origin. Hereditas 141:186–192.

Kolodinska Brantestam, A., R. von Bothmer, C. Dayteg, I. Rashal, S. Tuvesson, and J. Weibull. 2007. Genetic diversity changes and relationships in spring barley (Hordeum vulgare L.) germplasm of Nordic and Baltic areas as shown by SSR markers. Genet. Resour. Crop Evol. 54:749–758.

Kolodinska Brantestam, A., R. von Bothmer, I. Rashal, and J. Weibull. 2003. Changes in the genetic diversity of barley of Nordic and Baltic origin, studied by isozyme electrophoresis. Plant Genet. Resour. Charact. Util. 1:143–149.

Kruskal, J.B. 1977. Three-way arrays: Rank and uniqueness of trilinear decompositions, with applications to arithmetic complexity and statistics. Linear Algebra and its Applications 18:95–138. doi:10.1016/0024–3795(77)90069–6

Mackay, M.C., and K. Street. 2004. Focused Identi cation of

Germplasm Strategy–FIGS. Cereals. p. 138–141. *In* C.K. Black, J.F. Panozzo, and G.J. Rebetzke (ed.) Proc. of the Australian Cereal Chemistry Conf., 54th, and the Wheat Breeders' Assembly, 11th, Canberra, ACT. 21–24 September 2004. Royal Australian Chemical Institute, Melbourne, Australia.

Marshall, D.R. 1989. Limitations to the use of germplasm collections. p. 105–120. *In* A.D.H. Brown, O.H. Frankel, D.R. Marshall, and T.J. Williams (ed.) The use of plant genetic resources. Cambridge Univ. Press, Cambridge, UK.

MATLAB. 2009. MATLAB version R2007b. Available at http://www.mathworks.com/ (accessed 2 Dec. 2009; veri ed 27 July 2010). MathWorks Inc., Natick, MA.

Peeters, J.P., H.G. Wilkes, and N.W. Galwey. 1990. The use of ecogeographic data in the exploitation of variation from gene banks. Theor. Appl. Genet. 80:110–112 doi:10.1007/BF00224023.

Peeters, J.P., and J.T. Williams. 1984. Towards better use of genebanks with special reference to information. Plant Genet. Resour. Newsl. 60:22–32.

PLS-Toolbox. 2009. PLS-Toolbox version 5. Available at http://www.eigenvector.com (veri ed 27 July 2010). Eigenvector Research Inc., Wenatchee, WA.

Prada, D. 2009. Molecular population genetics and agronomic alleles in seed banks: Searching for a needle in a haystack? J. Exp. Bot. 60:2541–2552 doi:10.1093/jxb/erp130.

SESTO Genebank Information and Management System. 2009. Available at http://sesto.nordgen.org/ (veri ed 27 July 2010).

Smilde, A.K. 1997. Comments on multilinear PLS. J. Chemom. 11:367–377.

Smilde, A., R. Bro, and P. Geladi. 2004. Multi-way Analysis, Applications in the Chemical Sciences. Wiley, Chichester, UK.

Spooner, D.M., S.H. Jansky, and R. Simon. 2009. Tests of taxonomic and biogeographic predictivity: Resistance to multiple disease and insect pests in wild relatives of cultivated potato. Crop Sci. 49:1367–1376.

Street, K., M. Mackay, O. Mitrofanova, J. Konopka, M. El Bouhssini, N. Kaul, and E. Zuev. 2008. Swimming in the genepool—a rational approach to exploiting large genetic resource collections. *In* R. Appels, R. Eastwood, E. Lagudah, P. Langridge, M. Mackay, L. McIntyre, and P. Sharp (ed.) Proc. Int. Wheat Genetics Symp., 11th, Brisbane, Australia. 24–29 Aug. 2008. Sydney Univ. Press, Sydney, Australia.

Tanksley, S.D., and S.R. McCouch. 1997. Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. Science 277(5329):1063–1066 doi:10.1126/science.277.5329.1063.

Thuiller, W., J. Vayreda, J. Pino, S. Sabate, S. Lavorel, and C. Gracia. 2003. Large-scale environmental correlates of forest tree distributions in Catalonia (NE Spain). Glob. Ecol. Biogeogr. 12:313–325 doi:10.1046/j.1466–822X.2003.00033.x.

Tucker, L.R. 1966. Some mathematical notes on three-mode factor analysis. Psychometrika 31:279–311.

Varmuza, K., and P. Filzmoser. 2009. Introduction to Multivariate Statistical Analysis in Chemometrics. p. 35–36. CRC Press, New York.

Wold, S., M. Sjöström, and L. Eriksson. 2001. PLS-regression: A basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58:109–130.

Wold, H. 1966. Estimation of principal components and related models by iterative least squares. p. 391–420. *In* P.R. Krishnaiaah (ed.) Multivariate Analysis. Academic Press, New York.

van de Wouw, M., T. van Hintum, C. Kik, R. van Treuren, and B. Visser. 2010. Genetic diversity trends in twentieth century crop cultivars: A meta analysis. Theor. Appl. Genet. 120:1241–1252

van Hintum, Th.J.L., A.H.D. Brown, C. Spillane, and T. Hodgkin. 2000. Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome. ISBN: 9789290434542

Paper III

Endresen, Dag Terje Filip, Kenneth Street, Michael Mackay, Abdallah Bari, and Eddy De Pauw (2011). Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces. *Submitted to Crop Science on 21 December 2010. Conditionally accepted on 6 February 2011, first revision submitted 9 March 2011, accepted on 10 April 2011.*

Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces

Dag Terje Filip Endresen (1)*, Kenneth Street (2), Michael Mackay (3), Abdallah Bari (2), Eddy De Pauw (2)

(1) Nordic Genetic Resources Center (NordGen), Alnarp, Sweden. (2) International Center for Agricultural Research in the Dry Areas (ICARDA), Aleppo, Syria. (3) Bioversity International, Rome, Italy. * Corresponding author: dag.endresen@nordgen.org

ABSTRACT

Collections of crop genetic resources are a valuable source of new genetic variation for economically important traits, including resistance to crop diseases. New sources of useful crop traits are often identified through evaluation in field trials. The number of relevant accessions in genebank collections available to be evaluated is often substantially larger than the capacity of the evaluation project. The Focused Identification of Germplasm Strategy (FIGS) is an approach used to select subsets of germplasm from genetic resource collections in such a way as to maximize the likelihood of capturing a specific trait. This strategy uses a range of methods to link the expression of a specific trait (of a target crop) with the ecogeographic parameters of the original collection site. This study contributes to the development of the approach by which a FIGS subset could be assembled for biotic traits. We have evaluated trait specific subset selection methods for two fungal crop diseases namely stem rust (*Puccinia graminis* Pers.) in wheat (*Triticum aestivum* L. and *Triticum turgidum* L.) and net blotch (*Pyrenophora teres* Drechs.) in barley (*Hordeum vulgare* L.). The results indicate that the climate layers from freely available ecogeographic databases are well suited to model and predict the reaction in these crops to biotic stress traits. This result has the potential to improve the efficiency of field screening trials to find novel sources of economically valuable crop traits.

Abbreviations: BIOCLIM, bio-climatic; CI, confidence interval; FIGS, focused identification of germplasm strategy; FN, false negatives; FP, false positives; GIS, geographic information systems; kNN, k-nearest neighbor; LDA, linear discriminant analysis; LR+, positive diagnostic likelihood ratio; OR, odds ratio; PCA, principal component analysis; PET, potential evapotranspiration; PLS, partial least squares; PLS-DA, partial least squares discriminant analysis; PA, proportion positive agreement; PO, proportion observed agreement; PPV, positive predictive value; SIMCA, soft independent modeling by class analogy; TN, true negatives; TP, true positives;

INTRODUCTION

The genetic resources conserved by *ex situ* genebanks around the world cover a vast range of genetic diversity, underexploited in present day cultivars. The main objective of public genebanks is to conserve crop genetic diversity, to sustain agricultural production systems, by providing ready access to samples for research and plant breeding activities. A bottleneck for rational utilization is the availability and access to passport, characterization and evaluation

data. Obtaining good quality phenotypic trait data for genebank accessions requires large field or greenhouse experiments at great cost. The lack of evaluation data for useful traits is one of the major, current problems hindering the efficient use of plant genetic resources (FAO, 2010).

Further, the growing size of the genebank collections has been mentioned as a problem for the efficient use of genebank collections (see for example Mackay, 1990 and 1995) because the number of genebank accessions available to be evaluated for a specific trait is often substantially larger than the resources available to evaluate the material. Thus, finding the genebank accessions most likely to possess the desired trait can be compared to searching for a needle in a haystack. Clearly a rational and efficient strategy to mine genebanks for useful traits is required.

Focused Identification of Germplasm Strategy (FIGS)

The challenges to using genetic resource collections, as detailed above, was one of the reasons for the introduction of the core collection concept (Frankel, 1984; van Hintum et al., 2000). A core collection seeks to represent most of the genetic variation present in the original collection in a 'core' subset of 5-10% the size of the original. Core collection methods use statistical approaches to maximize diversity using a variety of input data including collection site descriptors, agro-morphological traits and molecular marker data.

However, the core collection approach may not lead to the identification of rare useful traits in germplasm collections (Holbrook and Dong, 2005). Such concerns to capture rare traits and adaptive trait variation, much of which thought to reflect plant functional variation (Wright and Gaut, 2005) have lead some workers to construct specific or thematic collections or use of other approaches (Brown and Spillane, 1999; Gepts 2006; Dwivedi et al. 2007; Pessoa-Filho et al., 2010; Xu, 2010)

The FIGS strategy introduces a novel approach for constructing small subsets of accessions in that it selects genetic variation for just a single trait at a time. The FIGS strategy endeavors to maximize the likelihood of encountering specific adaptive traits in subsets by choosing accessions from collection sites that are most likely to impose a selection pressure for the trait being sought. (Mackay and Street, 2004).

Nikolai Ivanovich Vavilov (1887-1943) was one of the first pioneers to recognize the importance of the ecoclimatic conditions when searching for source material to include in plant breeding (Vavilov, 1932; Vavilov, 1957; Dorofeyev, 1992; Kurlovich et al., 2000). Vavilov used the term 'climatic analogy' for the selection of suitable strains guided by climate and soil data. His 'differential phyto-geographical method' also has elements that link the morpho-physiological trait characters of species and strains to a definite environment and area (Vavilov, 1920, 1922, and 1935).

> *"It is evident that when selecting species and strains for the U.S.S.R. it is necessary to take the climate and the soil conditions at their origin into consideration in order to introduce strains from areas that are more or less similar to those in our own country. Knowledge of the climate of our own country and that of the areas from where we collected the seeds is of great importance."* (Vavilov, 1932; *cf*. Dorofeyev, 1992:266).

This link between environment and phenotype was recently demonstrated by Endresen (2010), who successfully used the FIGS strategy to link morphological traits in barley to the ecoclimatic pattern from the original collecting sites for Nordic barley landraces. Put into practice, a simple example of applying the FIGS approach to selection of germplasm from a genebank could be if salinity tolerance is the target trait then accessions would be chosen from collection sites that

have saline soils (Peeters et al., 1990). Hijmans et al. (2003) explored, based on a similar hypothesis, the link between frost tolerance and ecoclimate with focus on temperature at the original collecting site for genebank accessions.

However, the problem becomes more complex if one is looking for tolerances to biotic constraints. The approach used by El Bouhssini et al. (2009) to identify bread wheat resistance to Sunn pest (*Eurygaster integriceps* Puton) and to the virulent Syrian Russian wheat aphid biotype (*Diuraphis noxia* Kurdjumov) (El Bouhssini et al., 2011) was to select accessions from agro-climatic environments that are likely to favor high pest populations during the growing season. Thousands of accessions from the ICARDA (International Center for Agricultural Research in the Dry Areas) genebank had previously been chosen, largely at random, and screened for the two pests without success (Pers com., El Bouhssini). By contrast, the FIGS approach chose relatively small subsets (500 accessions) that contained multiple sources of resistance.

The first step when using the FIGS approach is to identify a group of georeferenced landraces with known resistance to a given pest or disease. An ecogeographic profile of the collection sites of this 'training set' is ascertained and statistical methods or models developed to select untested accessions from environments that are statistically similar to the 'trainer set' environments. This was the approach successfully used by Bhullar et al. (2009) and Bhullar et al. (2010) to identify a range of bread wheat accessions with resistance to various powdery mildew isolates.

The above examples demonstrate the utility of FIGS as a means to choose germplasm with variation for specific adaptive traits. However, they cannot be used as a proof of concept because the frequencies of the resistant material in the collection from which the subset was chosen were not known. Further, the possibility exists that the resistant material occurred in the sets merely by chance. The aim of this present study was to use geospatial statistical analysis to predict the presence of resistance to stem rust in bread wheat and net blotch in barley in a set of accessions that have been previously screened for the diseases, thus allowing an evaluation of the approach by comparing the predictions to a random selection.

Stem rust

The stem rusts are caused by the fungus *Puccinia graminis* (Pers., 1794) and are a significant disease affecting cereal crops. The *formae speciales* of *Puccinia graminis* f. sp. *tritici* is responsible for stem rust of wheat (McIntosh et al., 1995). E. C. Stakman provided early pioneering work on stem rust and identified the first unique races of this pathogen (Stakman, 1915; Stakman and Piemeisel, 1917). After a number of devastating rust epidemics not least in Australia, Canada and the USA, a long-term global collaboration to combat wheat rust was so successful that the stem rust reached almost non-significant levels by the 1990s (Singh et al., 2008). In 1998 a new isolate of stem rust designated Ug99 and typed to race TTKS caused severe damage to wheat in Uganda and Kenya. Pretorius et al. (2000) discovered that this new stem rust race Ug99 showed virulence against the widely used Sr31 stem rust resistance gene in wheat. This grasped again the full attention of crop scientists, and revived targeted international crop research collaboration against stem rust in 2005 with the Borlaug Global Rust Initiative (http://www.globalrust.org). When Ug99 continued to spread north through the Eastern African highlands and across the Red Sea into Yemen, it also reached the wider public through the media (see for example Koerner, 2010). Flood (2010) is warning that action to combat plant health problems in general, with a specific and concerted action to combat the Ug99 stem rust epidemic, is of vital importance to ensure food security. Identification of novel sources of resistance to stem rust is urgent and a current priority of a number of crop research groups around the world.

Net blotch

Net blotch is caused by a fungal pathogen (*Pyrenophora teres* Drechs.). It is known to cause serious harm to barley (*Hordeum vulgare* L.) and typically reduce yields by 10-40% (Steffenson, 1997). The disease thrives most under wet (high relative humidity) and warm conditions with temperature optima between 15-25° C depending on the region (Krupinsky et al., 2002). There are two common forms of the net blotch fungus. *Pyrenophora teres* f. *teres* produces a net like pattern, while *Pyrenophora teres* f. *maculata* produces more spot like lesions on the leaves of the crop plants (Liu and Friesen, 2010). Afanasenko et al. (1995) reported that the resistance against *Pyrenophora teres* f. *teres* (net type) and the *Pyrenophora teres* f. *maculata* (spot type) of net blotch are independently inherited. New epidemics of the net blotch have been reported around the world recently (McLean et al., 2009; Liu et al., 2010). Development of resistant varieties is the most cost-efficient method for control of net blotch (Jalli, 2010a; Jalli, 2010b). Silvar et al. (2010) reported finding few examples of resistance to net blotch in the Spanish barley core collection.

MATERIALS AND METHODS

This study used the results of disease screenings for landrace accessions maintained by the USDA NPGS GRIN (United States Department of Agriculture, National Plant Germplasm System, Germplasm Resources Information Network, http://www.ars-grin.gov/npgs/). Only those accessions with georeferenced collection sites were used in the study. Agro-climatic data, obtained from ICARDA (http://www.icarda.org) and WorldClim (http://www.worldclim.org), describing the collection sites were used to develop models to predict the presence of resistant phenotypes.

The stem rust trait dataset

The stem rust data is available online from the USDA NPGS GRIN database (http://www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?65049). Bonman et al. (2007) described the experimental design for the field trials. Susceptibility to stem rust (*Puccinia graminis* Pers. f.sp. *tritici*) was measured for six different years (during 1988-1994) at the agricultural research stations at St Paul (44º59'17" N, 93º10'48 W) and Rosemount (44º43'01" N, 93º05'56" W) located in Minnesota in the northern USA. Dr Don V. McVey made all of the trait observations for both locations. The trial experiments at Rosemont were inoculated by race TNMK. The trial experiments at St Paul were inoculated by race QFBS, RKQS and RTQQ. The trials at St Paul in 1988 and 1989 were also inoculated by race QSHS and RHRS; since 1991 also by HNLQ; and since 1992 also inoculated by race TNMK. The dataset contains observations for bread wheat (*Triticum aestivum* L.) and durum wheat (*Triticum turgidum* L.) and a total of 10 different subspecies. The original source locations for the wheat landraces are widely distributed across countries in Europe, Asia and Northern Africa (Figure 1). Ethiopia and Turkey are the best represented countries with each having more than 20% of the landraces from the stem rust dataset. Complementary georeferencing was made at ICARDA for landraces with missing geographic coordinates based on the description of the original collecting site. Only the 4932 landraces successfully georeferenced were included in this study. The original stem rust ratings (6889 trait observations) were reported as classified into ten classes according to the degree of reaction to the disease. The stem-rust trait ratings 0-3 (1915 landraces, 28% of the total) were considered as resistant to stem rust, ratings 4-6 (2729 landraces, 40%) as intermediate, and ratings 7-9 (2245 landraces, 32%) as susceptible. The complete stem rust dataset for this study including the ecoclimatic data is included as supplementary material (S2).

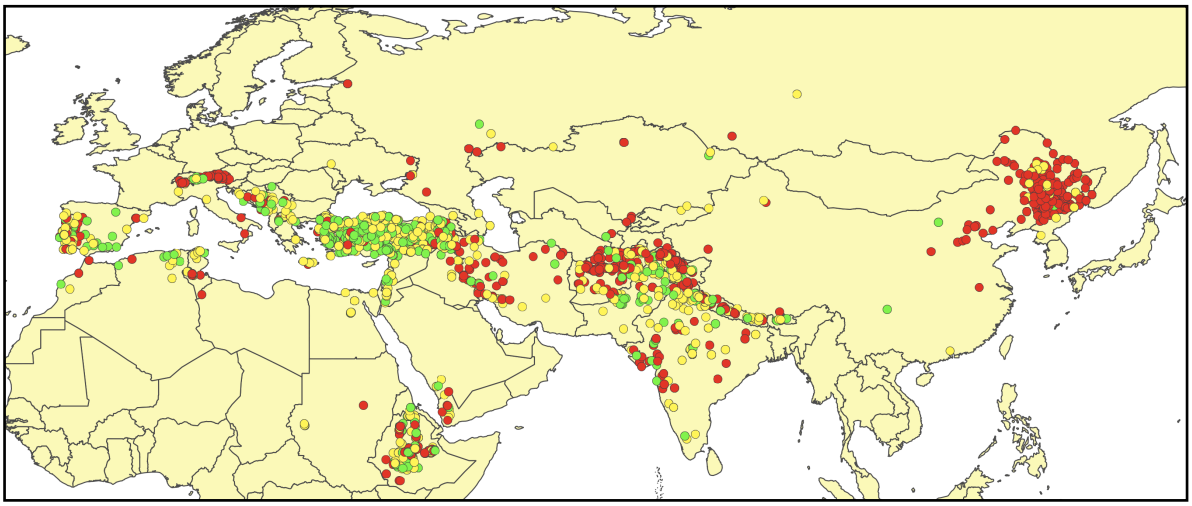


Figure 1: Distribution of the original source locations for the wheat landraces (stem rust dataset, 4932 accessions from 2013 collecting sites). Green circles indicate the collecting site for landraces resistant to stem rust, yellow circles show the medium susceptible, and red circles show where the susceptible landraces were collected.

The net blotch trait dataset

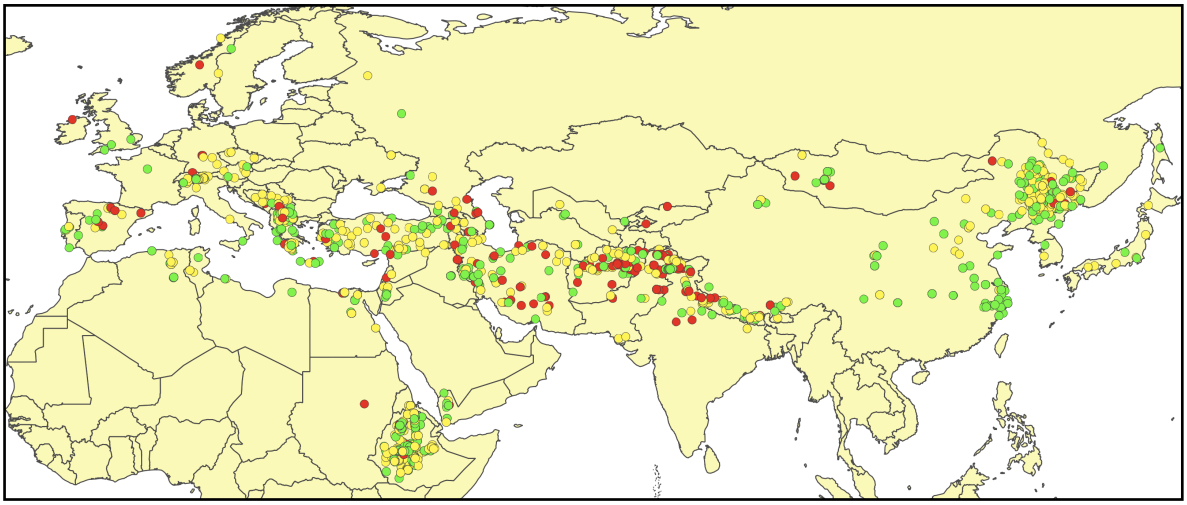


Figure 2: Distribution of the original source locations for the barley landraces (net blotch dataset, 2786 accessions). Green circles indicate the collecting site for landraces resistant to net blotch, yellow circles show the medium susceptible, and red circles show the locations where susceptible landraces were collected.

The net blotch trait dataset is available online from the USDA NPGS GRIN database (http://www.ars-grin.gov/cgi-bin/npgs/html/desc.pl?1041). The FIGS net blotch set was extracted from the USDA GRIN NPGS database by Dr Harold Bockelman and includes trait observations for the reaction to net blotch (*Pyrenophora teres* Drechs.) for a total of 4645 barley (*Hordeum vulgare* L.) landraces (including greenhouse observations). Although net blotch seedling data from the greenhouse were available, they did not differentiate the landraces to the degree of the field trials. Thus, for this study we decided to focus on the analysis of the trait ratings from the field trials. From the net blotch dataset a total of 2786 georeferenced accessions were tested under field conditions during eight different years (1988-2004) at four different agricultural research stations, Langdon, ND (48º45'43" N, 98º22'20" W), Stephen, MN (48º27'03" N, 96º52'30" W), Fargo, ND (46º52'37" N, 96º47'20" W), and Athens, GA (33º57'18" N, 83º22'59" W). The field trial experiments were inoculated by isolate ND89-19 of net blotch

(Pyrenophora teres f. teres) using infected barley straws from the previous season (Bonman et al., 2005). The original net blotch trait ratings (2786 trait observations) were reported as classified into nine classes with ratings 1-3 (1115 landraces, 40% of total records) considered as resistant to net blotch, ratings 4-6 (1367 landraces, 49%) as intermediate, and ratings 7-9 (304 landraces, 11%) as susceptible. The original source locations for the barley landraces are widely distributed across 51 countries in Asia, Europe and Northern Africa (Figure 2). A total of 1025 (36.8%) of the landraces originate from Ethiopia. The next country ranked by total number of records in the dataset was China with 365 (13.1%) landraces. The complete net blotch dataset for this study, including climate data, is included as supplementary material (S3).

ICARDA ecoclimatic database

The ICARDA ecoclimatic information system was created in 2003 covering the CWANA region (Central and West Asia and North Africa) (De Pauw, 2008). In 2005 the dataset was further extended to cover Europe and most of Asia. The 'thin-plate smoothing spline' method of Hutchinson (1995), as implemented in the ANUSPLIN software (Hutchinson, 2000), was used to convert the station based climatic database into 'climate surfaces' with a 30 arc-second (approximately 1 km) resolution grid. The dataset includes monthly mean values for minimum temperature (tmin), maximum temperature (tmax), precipitation (prec), and potential evapotranspiration (pet), as well as a wide range of derived climatic variables, such as agroclimatic zone, aridity index, length-of-growing-period, and others. The soil layers from the ICARDA ecoclimatic information system are originally derived from the FAO Soil Map of the World (FAO, 1974; FAO-UNESCO, 1995; FAO, 2007). The climate data for this study was extracted for each accession using the longitude and latitude coordinates of the original collecting site.

WorldClim ecoclimatic database

Climate data for the net blotch dataset was extracted from the WorldClim dataset (http://www.worldclim.org; Hijmans et al., 2005) with DIVA GIS (http://www.diva-gis.org). The WorldClim dataset was developed (with the ANUSPLINE software) following a similar method as described above for the ICARDA ecoclimatic information system. The climatic layers from the WorldClim dataset include monthly mean values for temperature (temp), minimum temperature (tmin), maximum temperature (tmax), precipitation (prec), and the derived BIOCLIM layers (Busby, 1991). The WorldClim dataset is available in different spatial resolutions: 30 arc-seconds (approximately 1 km), 2.5 minutes (approximately 4.5 km = 20 km^2), 5 minutes (approximately 9.3 km = 86 km^2), and 10 minutes (approximately 18.5 km = 342 km^2).

Trait mining models

The underlying hypothesis for this study was that certain types of environments would favor the emergence of disease resistance within *in situ* populations of landraces. Negri et al. (2009:9) proposed to define a landrace as a cultivated plant with the *"lack of formal crop improvement"* and *"characterized by a specific adapation to the environmental conditions of the area of cultivation"*. The general approach was to classify collection site environments into those that are likely to yield a certain category of disease reaction. This was achieved by applying classification models to one set of data in which the disease score is used to 'train' the model so that it correctly classifies environments according to disease type. Another "test" subset was used to proof the classification model using the disease scores for the selected accessions.

Validation subsets (training set + test set)

The training-set used to calibrate the prediction models comprised 67% of the records, and the test set made up 33% of the records. For each subset the allocation of samples to the training set

and the test set was a random process. The calibration of the model parameters was made with cross-validation (Hawkins et al., 2003). The samples from each of the test sets were only used to validate the predictive performance and, were not included in any calibration steps (Hawkins, 2004).

Classification algorithms

The predictive performance for four different classification methods was compared.

(1) Linear discriminant analysis (LDA). The LDA (Fisher, 1936) classifier assumes a normal (Gaussian, bell curve) distribution for the predictor variables. Also the residuals are assumed to show the normal distribution. The normality assumption can be tested for example with a Q-Q plot or a Wilk-Shapiro test (Shapiro and Wilk, 1965). The LDA method requires more assumptions to be met than the other classification methods of this study. However parametric methods usually outperform nonparametric methods when the assumptions to the underlying data distribution are met. It would thus always be wise to include a parametric classification method when the assumption of normality is evaluated to be reasonable.

(2) Partial Least Squares Discriminant Analysis (PLS-DA). PLS-DA (Barker and Rayens, 2003) is based on the Partial Least Squares (PLS) (Wold, 1966; Wold et al., 1984). With PLS-DA the training set is used to calibrate new latent variables that can be seen as a linear combination of the previous multivariate variables. The unknown samples are projected into this new multivariate space defined by the latent variables. A separate sub-model is calibrated for each class. All of these sub-models are calibrated together in an iterative process to simultaneously fit the independent, predictor variables (climate data) and the dependent, response variables (trait data). Compression of the original multivariate variables to latent factors or principal components will often provide a solution to the common problem of co-linearity between the predictor variables.

(3) k-nearest neighbor algorithm (kNN). kNN is a pattern recognition method for classification of objects based on the majority vote of the closest neighbors (Cover and Hart, 1967). This is one of the simplest classification algorithms. The kNN algorithm is a non-parametric classification method, and thus makes no assumption on the underlying data distribution (Duda et al., 2001). For all the kNN classifiers in this study we used k=1, where the observation is assigned to the class of its immediate/nearest neighbor.

(4) Soft independent method of class analogy (SIMCA). SIMCA is a method of classification where separate PCA (principal component analysis) models are calibrated for each class in the dataset (Wold and Sjostrom, 1977; Wold, 1976) (similar as for PLS-DA). These sub-models are calibrated independently, and only fitted to the independent predictor variables (different from the PLS-DA algorithm). The unknown samples are fitted to each of the sub-models and assigned to the sub-model of the closest fit. An unknown sample can in practice be assigned to multiple classes or to no class. The SIMCA method is a projection method where the multivariate dataset under study is condensed to a set of lower-dimensional subspaces (PCA models). The SIMCA method requires few other assumptions than that the condensed subspaces provide a meaningful representation of the original dataset.

(*) Proofing the performance of the classification algorithms against a random selection. For each test set the classification results were compared to the results from an actual random selection. For the random selection, the trait scores for each test set where reassigned using random permutation of accession numbers. Using this approach the number of examples for each trait category remained unchanged, but any link between the trait and the ecoclimatic

description was broken. With the random selection experiments we can directly compare the performance of the classification methods to the effect of a random selection in practice. The behavior and stability of the performance indicators compared to a random selection are illustrated here with a practical test.

All of the classification tests for each subset were repeated 3-5 times to provide a series of replicated classification experiments. For each of these series the subset was split again by random into a new training set and a new test set. The classification indicators reported below are the average from these replicated classification experiments.

Trait mining pre-study

Before starting the trait mining experiments a series of pre-study tests were made to identify the most appropriate classification algorithm (choosing from kNN, SIMCA, PLS-DA, or LDA), and also the most appropriate number of levels for the trait score measurement scale (choosing from 2, 3, or 9 category levels). The most appropriate classification method and number of levels for the trait scale was next used for the corresponding series of trait mining experiments described below. The results from the pre-study tests are reported in Tables 1, 2, and 3.

Performance for the models when the stem-rust disease score categories were rescaled. This study was designed to explore if reclassification to fewer trait categories might contribute to improved predictions for the resistant landraces. The objective of this study was not to predict accurate levels of disease susceptibility, but to identify the resistant landraces. The degree of disease susceptibility among the susceptible landraces was not the aim of this trait mining experiment. The disease score categories for stem rust were rescaled from a 0-9 scale as follows: **S2** = disease scores reclassified to 2 classes to represent resistance (0-3) and susceptibility (4-9). **S3** = disease scores reclassified to 3 classes (0-3), (4-6), (7-9). **S9** = disease scores reclassified to 9 classes (0-1), 2, 3, 4, 5, 6, 7, 8, 9. The net blotch dataset did not include the original trait score zero (0), but the reclassification followed in every other respect exactly the same schema as for the stem rust set.

Trait mining experiments

The modeling experiments described here explore the predictive performance when using different data stratification strategies and different ecoclimatic data sets.

Experiment 1: *Performance of models when different ecoclimatic data is used.* In this experiment the predictive power of the models to identify stem rust resistance in wheat were compared using the 3 different ecoclimatic data sets: the ICARDA ecoclimatic dataset respectively with and without the layers for potential evapotranspiration included (ICARDA-30", and ICARDA-30"-w-pet; with 30 seconds = 1 km spatial resolution), the WorldClim dataset (WorldClim-2.5'; 2.5 minutes = 4.5 km resolution) and the BIOCLIM layers (Busby, 1991) derived from the WorldClim layers (BIOCLIM-2.5'; 4.5 km resolution). The results from experiment 1 are included in Table 4.

Experiment 2: *Performance of models after the data is stratified according to genetic background.* In this experiment the predictive power of the SIMCA models to identify stem rust resistance in wheat landraces were compared for two subsets containing data for *Triticum aestivum* L. ssp *aestivum* L., and *Triticum turgidum* L. ssp *durum* (Desf.) Husn. The working hypothesis was that the different taxa (genetic background/genome) of wheat might have slightly different mechanisms of resistance against stem rust and thus a different predictive association between the ecogeographic parameters and the trait score. The results from experiment 2 are included in Table 5.

Experiment 3: *Performance of models after the data is stratified according to test site.* In this experiment the predictive power of the SIMCA models to identify stem rust resistance in wheat were compared for two subsets. One subset containing data for scores obtained in St Paul, USA and the other containing scores obtained in Rosemount, USA. The results from experiment 3 are included in Table 5.

Experiment 4: *Performance of models after the data was stratified according to year of screening.* In this experiment the predictive power of the SIMCA models to identify stem rust resistance in wheat was compared for 6 subsets corresponding to what year the disease scoring was undertaken: 1988, 1989, 1991, 1992, 1993 or 1994. The results from experiment 4 are included in Table 5.

Note that for experiment 1, 2, 3, and 4 collection site ecoclimatic data were included for all accessions scored. A total of 6889 records with stem rust ratings (corresponding to a total of 4932 genebank accessions) were processed, meaning that some of the sites (2013 sites in total) were represented in the analysis multiple times.

Experiment 5: *Performance of the models when only one accession per site is included in the analysis.* The complete trait datasets from the USDA includes only very few multiple measurements for the same landrace (replications), but there are often multiple landraces in the dataset that originate from the same source location. Thus, the ecogeographic parameters for the landraces collected at the same source location are identical. A total of 1124 landraces from the Ethiopian stem rust set originated from only 191 different source locations. For India, Afghanistan and Turkey there were also more than twice as many landraces as source locations. This trait mining experiment in the stem rust set was designed to explore the effect of pooling observations by distinct source locations. The hypothesis was that source locations with high sample number would tend to dominate the models, and that reducing the number of landraces to one per collecting site would give the ecogeography of each site more equal influence on the prediction. In experiment 5 the modeling design for experiment 1 above was repeated using only one accession per site. Thus only 2013 records were processed in total for each experiment. The disease scores were averaged across accessions for sites containing multiple entries, thus only one (average) disease rating represented sites with multiple accessions. The results from experiment 5 are included in Table 6.

Experiment 6: *Performance of the models to select net blotch resistant barley accessions when using different ecoclimatic datasets.* In this experiment the predictive power of the models to identify net blotch resistance in barley were compared using the 3 different ecoclimatic datasets: the ICARDA ecoclimatic dataset with and without the layers for potential evapotranspiration (ICARDA-30", and ICARDA-30"-w-pet; with 30 seconds = 1 km spatial resolution), the WorldClim dataset (WorldClim-2.5' = 4.5 km resolution, WorldClim-5' = 9.3 km, and WorldClim-10' = 18.5 km) and the BIOCLIM layers (Busby, 1991) derived from the WorldClim layers (BIOCLIM-2.5', BIOCLIM-5', and BIOCLIM-10'). The results from experiment 6 are included in Table 7.

Evaluation of the trait mining results

The target of this trait mining study was concerned with the identification of resistant landraces rather than with the accurate classification of the samples to the different stem rust score categories. Thus, to assess the positive predictive performance for the identification of resistant landraces, the so-called confusion matrices (Kohavi and Provost, 1998) for each trait-mining experiment was first collapsed to a 2 by 2 table. The collapsed confusion matrix tabulated the

predicted resistant and susceptible landraces against the actual observed number of resistant and susceptible samples from the test sets. The samples, scored as 0-3 for the stem rust set and 1-3 for net blotch (on the original trait scale) were pooled together and classified as resistant samples. The samples scored as 4-9 on the original trait scale were classified as susceptible samples. The landraces predicted to be resistant were either true positive (TP) when a landrace observed to be resistant was also predicted to be resistant (positive); or they were false positives (FP). Like-wise the landraces predicted to be susceptible were either true negatives (TN), or false negatives (FN).

To establish and compare the predictive performance for the different test models, we need metrics to compare the models. A number of different indicators of inter-rater agreement have been developed for different purposes (Gwet, 2010). When choosing the indicator to use it is important to remember the aim of the current study was to identify a smaller subset of landraces more likely to be resistant to stem rust than what would be expected when selecting landraces by chance. We are thus more concerned with the specificity of the model to identify the resistant samples than on the overall agreement related to correctly predicted samples.

Cohen's Kappa (Kappa) is a popular measure of inter-rater agreement for categorical (qualitative) items (Cohen, 1960). Cohen's Kappa is applied here as an indicator for the agreement between the actual observed trait scores and the predicted trait category as calculated by the models. The Kappa inter-rater agreement indicator aims to calculate the observed agreement adjusted for the level of agreement that would be expected by chance. The value of Cohen's Kappa ranges from -1 to +1, where +1 implies perfect agreement, 0 implies no relationship and -1 implies perfect disagreement (Landis and Koch, 1977). There remains some controversy in particular related to the application and calculation of agreement expected by chance. Cohen's Kappa is however widely used in the absence of standard, general accepted alternatives (Gwet, 2010).

Among the other alternative indicators for predictive performance the raw proportion observed agreement (**PO**) is perhaps the most intuitive (Altman and Bland, 1994a). The total number of overall observed agreement (TP + TN) is simply divided by the total number of samples (N). The proportion positive agreement (**PA** = 2*TP / (2*TP + FP + FN)) might be a more appropriate indicator for our aim to identify the resistant (positive) samples. The positive predictive value (**PPV** = TP / (TP + FP)) is yet another suitable indicator to measure the predictive performance for identification of resistant landraces (Altman and Bland, 1994b). The PPV measures the probability that a landrace predicted by the model to be resistant is truly resistant (observed as resistant in the test set). The PPV is inherently dependent on the prevalence of resistant samples. However most of the trait mining test sets for this study have the same proportion of resistant samples as the overall trait dataset, making the PPV a suitable indicator to compare the performance of these modeling experiments. Some subsets (e.g. split by experiment year or location) show very different prevalence for the proportion resistant samples. The prevalence was evaluated as the number of resistant samples in relation to all samples for each test set. The positive diagnostic likelihood ratio (**LR+** = (TP / (TP + FN)) / (FP / (FP + TN)) = sensitivity / (1 - specificity)) is more appropriate when comparing test sets with very different prevalences (Altman and Bland, 1994b). The positive likelihood ratio (LR+) measures how much more likely it is for the model to predict a landrace to be resistant (positive) in the group of landraces observed to be resistant compared to making this prediction in the group of landraces observed to be susceptible. For each of the modeling tests we have also calculated the sensitivity (**Sensitivity** = TP / (TP + FN)) and the specificity (**Specificity** = TN / (TP + FP)) indicators (Altman and Bland, 1994a). The sensitivity is the proportion of actual resistant landraces that are correctly identified (predicted by the model) as such. The specificity

is the proportion of susceptible landraces that are correctly identified. The last performance indicator reported here is the Yule's Q (**Yule's Q** = (OR -1) / (OR +1); where the odds ratio, OR = (TP * TN) / (FN * FP)). The odds ratio can be interpreted as the magnitude of association between the model predictions and the actual observed trait values. The Yule's Q is only a transformation of the odds ratio so that the indicator takes values in the range between -1 and 1. For more information and additional other alternative indicators for the predictive performance for classification tests see the text book by Agresti (2002) or the more recent text book by Gwet (2010).

All of the reported performance metrics are the average from a series of replicated trait mining tests for each subset. The estimate for the gain or improved predictive performance compared to a random selection for each trait-mining test set is included in the right-side column of the results tables as the **Estimated Gain**. The Gain is calculated as the Positive Predictive Value (PPV) divided by the proportion resistant samples (prevalence) for each subset. The prevalence was calculated as the proportion resistant samples for each test set.

To further illustrate in practice the performance for a complete random selection, a series of actual random selections were made for each test set (with the same number of replications as for the trait mining models). The average of the tests with the random selection is reported at the bottom of the tables (rows marked "**Random selection**").

For this study all performance indicators are reported in the supplementary material (S1). For clarity however, in the results section only the PPV, LR+, and the Estimated Gain indicators are reported.

The 95% confidence interval is reported for the performance indicators. The confidence intervals are included in a parenthesis after each indicator. The reported confidence interval upper and lower boundaries were calculated with the online Statistics Calculator from the Centre for Evidence Based Medicine (CEBM, http://ktclearinghouse.ca/cebm/practise/ca/calculators/statscalc).

Software

The classification models were calculated with the MATLAB software and the PLS Toolbox (version 5.8, http://software.eigenvector.com/toolbox/). The Cohen's Kappa indicator was calculated with a MATLAB script by Cardillo (2007).

RESULTS

The initial explorative Principal Component Analysis (PCA) showed no clear grouping for the trait score categories (for either the stem rust or the net blotch sets). Only one sample was identified as a very strong outlier. This was landrace PI 212925 from the stem rust set. We could not identify why this landrace behaves as an outlier, yet the sample was very atypical of the other samples. We decided to remove this sample from the further analysis as reported below (6889 records with trait observations for 4932 wheat accessions collected from 2013 distinct sites). The initial PCA analysis did not identify any notable outliers in the net blotch set (2786 trait observations from the same number of barley accessions).

Pre-study 1

The purpose of the first pre-study (pre-study 1) reported in Table 1 was to identify the most appropriate classification algorithm to use in the experiments reported below, as well as the

most appropriate score reclassification scheme to use. All of the models used were able to significantly improve upon a random selection to capture stem rust resistance. For example, even the lowest performing classification model (PLS-DA) was 1.57 (Estimated Gain, Table 1) times more effective when compared to a random selection. However, there were significant differences in the performance of the classification models. The PPV, LR+ and Estimated Gain performance indicators reported in Table 1 demonstrate that the SIMCA model had the most effective predictive power; followed by the kNN, LDA and PLS-DA models, the latter of which yielded performance indicators that were significantly lower than the others. All the classification algorithms have significant higher predictive performance indicators than the random selection (not overlapping confidence interval).

Table 1: Comparison of the performance of 1) different classification models and 2) the performance of the SIMCA model using different levels of the trait measurement scale, to predict the occurrence of stem rust resistance in wheat (pre-study 1).

Dataset	Model	Scale	PPV	LR+	Estimated Gain
Stem rust	LDA	S3	0,48 (0,45-0,52)	2,40 (2,18-2,64)	1,73 (1,61-1,85)
Stem rust	PLS-DA	S3	0,44 (0,41-0,47)	2,01 (1,86-2,17)	1,57 (1,46-1,68)
Stem rust	kNN	S3	0,49 (0,45-0,53)	2,46 (2,17-2,77)	1,75 (1,61-1,88)
Stem rust	**SIMCA**	**S3**	**0,54** (0,50-0,59)	**3,07** (2,66-3,54)	**1,95** (1,79-2,09)
Stem rust	SIMCA	S9	0,53 (0,48-0,57)	2,86 (2,47-3,32)	1,88 (1,73-2,04)
Stem rust	SIMCA	S2	0,51 (0,48-0,55)	2,72 (2,42-3,07)	1,84 (1,70-1,97)
Random			*0,29 (0,26-0,33)*	*1,04 (0,90-1,20)*	*1,03 (0,91-1,16)*

The reported performance indicators are the average from 3-5 replications (with different random split to training set + test set) for the same full stem rust dataset. The record level data unit is here the trait measurement from the USDA net blotch set (6889 observations). The proportion true resistant samples in this dataset were 28% (prevalence = 0,28). The 95% confidence interval is included inside the parentheses. The positive predictive value (PPV) provides an indicator for classification performance of resistant samples (positives). The positive diagnostic likelihood (LR+) provides a similar indicator that is less sensitive to the prevalence or proportion of resistant samples (positives) in the dataset.

The degree to which the score data was re-categorized had an effect on the predictive power of all of the models. The results for the SIMCA model, detailed in Tables 1, 2, and 3 show that when the original 0-9 scoring categories are reclassified to 3 categories the models perform better than if asked to predict membership in 2 categories or 9 categories.

Pre-study 2

The second pre-study (pre-study 2) also examined the most appropriate classification algorithm to use as well as the most appropriate score reclassification scheme. This study differs from pre-study 1 in that the disease scores for sites represented by multiple accessions were averaged. Thus only one record per site was included in the analysis.

The trends demonstrated in Table 1 are corroborated by the results in Table 2. That is, the most effective classification algorithm to use is SIMCA, and 3 trait categories is the most appropriate trait scale to maximize the predictive performance of the models. However, processing the USDA stem rust dataset to include unique records for each distinct collecting site made a substantial improvement to the predictive performance. Note that the proportion of resistant cases (prevalence) is lower in the set with distinct sites (20% compared to 28% for the full set, see Table S1.1 of the supplementary material). The PPV indicator is sensitive to the prevalence,

so we should here rather focus on the more robust LR+ (and the estimated gain) than on the PPV indicator when comparing the performance for this dataset (pre-study 2) with the full set (pre-study 1). The estimated gain for the SIMCA model when using unique collection sites is 2.51 (150% higher hit rate compared to a random selection) compared to the gain of 1.95 (95%) when multiple accessions per site are processed. This is a substantial improvement for this indicator (even if the confidence intervals for the same classification algorithms in pre-study 1 and 2 are marginally overlapping). The LR+ indicator also shows a substantial improvement for the experiment with distinct collecting sites. In other words, these results indicate that it is better to average the scores across accessions when there are multiple accessions per site than to have a particular site represented multiple times in the analysis. In pre-study 2 (same as for pre-study 1) all of the classification algorithms have significant higher predictive performance indicators than the random selection (not overlapping confidence interval).

Table 2: Comparison of the ability of different classification models and different number of levels of the trait measurement scale, to predict the occurrence of stem rust resistance in wheat (pre-study 2).

Dataset	Model	Scale	PPV	LR+	Estimated Gain
Stem rust (site)	LDA	S3	0,39 (0,33-0,46)	2,60 (2,12-3,23)	1,97 (1,65-2,32)
Stem rust (site)	PLS-DA	S3	0,37 (0,31-0,44)	2,42 (2,00-2,93)	1,89 (1,57-2,20)
Stem rust (site)	kNN	S3	0,44 (0,36-0,53)	3,21 (2,44-4,28)	2,23 (1,82-2,65)
Stem rust (site)	**SIMCA**	**S3**	**0,50** (0,40-0,60	**4,00** (2,85-5,66)	**2,51** (2,02-2,98)
Stem rust (site)	SIMCA	S9	0,49 (0,41-0,59)	3,96 (2,91-5,43)	2,50 (2,03-2,93)
Stem rust (site)	SIMCA	S2	0,47 (0,39-0,55)	3,58 (2,75-4,69)	2,37 (1,96-2,77)
Random			*0,19 (0,13-0,26)*	*0,94 (0,63-1,39)*	*0,95 (0,66-1,33)*

The record level data unit is here the collecting site from the stem rust dataset (total 2013 sites, prevalence = 0.20). The disease trait scores for all wheat landraces collected at the same site were averaged to make one data record for each distinct collecting site. The 95% confidence interval is included inside the parenthesis.

Pre-study 3

The aim of the third pre-study (pre-study 3) was to determine the most effective classification model to predict the occurrence of net blotch resistance in barley, the results of which are detailed in Table 3. As for the pre-studies in the stem rust set (pre-study 1 and 2), the SIMCA model out-performed the other models. Similar trends in relative predictive powers of the models were also demonstrated (SIMCA > kNN > LDA > PLS-DA). In this study the performance of the SIMCA model was 35% higher than for an expected random selection.

What is notable though is that while the LR+ and estimated gains demonstrate that net blotch resistance can be effectively predicted using ecoclimatic parameters, the predicative power of the models is significantly lower than that demonstrated for the predictions in the stem rust set. For both LR+ and the estimated gain, the confidence intervals (95% CI) in the stem rust set are above and not overlapping with the confidence interval for these indicators in net blotch set. The 95% confidence interval using the SIMCA method was estimated to: LR+, stem rust set with distinct sites = (2.85-5.66); LR+, stem rust set = (2.66-3.54); LR+, net blotch set = (1.42-2.17). For pre-study 3 the PLS-DA algorithm have marginally overlapping confidence interval for the performance indicators when compared to the random selection. The other classification algorithms have significant higher predictive performance than the random selection, same as for pre-study 1 and 2.

Table 3: Comparison of the ability of different classification models, to predict the occurrence of net blotch resistance in barley (pre-study 3).

Dataset	Model	Scale	PPV	LR+	Estimated Gain
Net blotch	LDA	S3	0,50 (0,45-0,56)	1,52 (1,29-1,79)	1,26 (1,13-1,39)
Net blotch	PLS-DA	S3	0,48 (0,42-0,55)	1,41 (1,15-1,77)	1,21 (1,06-1,37)
Net blotch	kNN	S3	0,51 (0,46-0,56)	1,56 (1,33-1,82)	1,27 (1,14-1,40)
Net blotch	**SIMCA**	**S3**	**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)*

The record level data unit is here the trait measurement from the USDA net blotch set (2786 observations, prevalence = 0.40). The 95% confidence interval is included inside the parentheses.

Based on the above pre-studies (Table 1, 2, and 3), the trait mining experiments reported with Tables 4-7, the models were calibrated with the SIMCA classification algorithm using a trait measurement scale with 3 categories.

Experiment 1
The aim of this experiment was to see if there were significant differences between how the different ecoclimatic datasets affect the predictive power of the models. The results for the SIMCA model are reported in Table 4.

Table 4: Performance of the SIMCA model to select stem rust resistant accessions using different ecoclimatic layers (ICARDA, WorldClim, and BIOCLIM) compared to a random selection (experiment 1).

Ecoclimate	PPV	LR+	Estimated Gain
ICARDA 30" w-pet	0,55 (0,51-0,59)	3,08 (2,67-3,54)	1,94 (1,81-2,11)
ICARDA 30"	0,57 (0,52-0,61)	3,31 (2,87-3,82)	2,00 (1,87-2,17)
WorldClim 2.5'	**0,57** (0,52-0,61)	**3,54** (3,05-4,10)	**2,10** (1,94-2,25)
BIOCLIM 2.5'	0,49 (0,45-0,53)	2,48 (2,16-2,86)	1,76 (1,59-1,89)
Random selection	*0,29 (0,26-0,33)*	*1,04 (0,90-1,20)*	*1,03 (0,91-1,16)*

The record level data unit is here the trait measurement from the USDA stem rust set (6889 observations, prevalence = 0.28). The 95% confidence interval is included inside the parentheses.

While the performance indicators for the SIMCA model all score slightly higher for the WorldClim 2.5' dataset than the two ICARDA sets, the confidence intervals indicate that there is little difference between these sets (Table 4). This result was expected because both the WorldClim and the ICARDA ecological database are constructed using the same spatial interpolation method and is to a large extent (although not fully) based on the same climatic data sources. However, the model's predictive power was significantly lowered when the BIOCLIM-2.5' set (which are a different class of ecoclimatic parameters) was used. Despite this, using the BIOCLIM-2.5' set of ecoclimatic parameters was still 73% more effective than making a random selection. The BIOCLIM set is derived from the WorldClim set, and these results thus indicate that using the 'raw' WorldClim layers give a better predictive performance for this dataset than using the derived BIOCLIM layers.

Experiment 2

The aim of this study was to see if there was a difference in the predictive performance of the SIMCA model when applied to discrete datasets for *Triticum aestivum* ssp *aestivum* and *Triticum turgidum* ssp *durum*.

Table 5: Performance of the SIMCA model to select stem rust resistant accessions, after stratifying the trait data according to, 1) species 2) location of trial 3) year of trail, compared to a random selection (experiment 2, 3, and 4).

Stratified subset	LR+	Estimated Gain	Prevalence[†]
Bread wheat	**4,27** (3,42-5,28)	**2,57** (2,28-2,90)	0,20
Durum wheat	1,76 (1,43-2,17)	1,32 (1,17-1,45)	0,44
St Paul, USA	**3,07** (2,62-3,59)	**2,13** (1,93-2,40)	0,21
Rosemount, USA	2,54 (2,01-3,21)	1,55 (1,40-1,71)	0,41
Trial year 1988	4,44 (3,23-6,12)	**3,70** (2,37-5,28)	0,06
Trial year 1989	2,16 (1,82-2,55)	1,38 (1,28-1,49)	0,48
Trial year 1991	2,53 (2,04-3,12)	1,92 (1,58-2,21)	0,21
Trial year 1992	4,67 (3,19-6,82)	2,28 (1,87-2,57)	0,29
Trial year 1993	**6,18** (3,35-11,40)	2,17 (1,74-2,44)	0,36
Trial year 1994	2,86 (1,09-7,51)	2,18 (0,89-3,80)	0,17
Random selection	*1,30 (1,02-1,68)*	*1,21 (0,98-1,47)*	*0,26*

The stem rust dataset is here, stratified by the St. Paul experiment station and the Rosemount experiment station, both located in the northern USA; by 6 distinct trial years 1988, 1989, 1991, 1992, 1993, and 1994; and by the two wheat subspecies with the most data records (landraces). The record level data unit is here the original observation measurement from the USDA stem rust set. The 95% confidence interval is included inside the parentheses.

[†]The PPV indicator is not included to this table because the PPV is sensitive to the prevalence, which is very variable for these stratified subsets.

When the dataset was stratified based on genetic background, the LR+ indicator showed an even higher response; 4.27 for bread wheat compared to 1.76 for durum wheat. The bread wheat subset has substantially higher predictive performance compared to the random selection. The durum wheat subset has overlapping confidence interval (95% CI) with the random selection and we have thus no significant predictive effect from the models in this subset.

Experiment 3

The predictive performance of the SIMCA model also differed when the stem rust data was split according to disease test site. The models performance using the St Paul data was significantly higher than those obtained for Rosemount (LR+ = 3.07 and 2.54 respectively, Table 5). Both subsets (St Paul and Rosemount) however performed significantly better when compared to a random selection.

Experiment 4

Stratifying the data according to year of trial also significantly affected the performance of the models. Again, notice that the stratified subsets often have very different proportions of resistant samples (prevalence), so the positive diagnostic likelihood ratio (LR+) should be compared

rather than the Positive Predictive values (PPV). Despite the differences in the models' performance when the data were split into years of experiment each result was significantly higher than a random selection process (LR+, Table 5).

Experiment 5
Processing the USDA stem rust dataset to include unique records for each distinct collecting site made a substantial improvement to the predictive performance in that the magnitude of the LR+ and Estimated Gain was greater in this experiment (Table 6) than those reported for experiment 1 (Table 4). Note that the proportion resistant cases are lower in the set with distinct sites (20% compared to 28% for the full set).

Table 6: Comparison of the performance of the SIMCA model to select wheat accessions resistant to stem rust using different ecoclimatic data (ICARDA, WorldClim, BIOCLIM) and with distinct collecting sites as the record level data unit (experiment 5).

Ecoclimate	PPV	LR+	Estimated Gain
ICARDA 30'' w-pet	0,53 (0,44-0,65)	**4,70** (3,41-7,47)	**2,76** (2,29-3,39)
ICARDA 30''	0,43 (0,33-0,53)	3,16 (2,13-4,51)	2,23 (1,73-2,74)
WorldClim 2.5'	**0,55** (0,46-0,67)	4,46 (3,20-7,02)	2,55 (2,09-3,09)
BIOCLIM 2.5'	0,46 (0,40-0,56)	3,55 (2,93-4,96)	2,38 (2,08-2,90)
Random selection	*0,19 (0,13-0,26)*	*0,94 (0,63-1,39)*	*0,95 (0,66-1,33)*

The record level data unit is here the collecting site from the stem rust dataset (total 2013 sites, prevalence = 0.20). The disease trait scores for all wheat landraces collected at the same site were averaged to make one data record for each distinct collecting site. The 95% confidence interval is included inside the parentheses. ICARDA-30''-w-pet includes monthly evapotranspiration.

In experiment 1 (Table 4) the WorldClim-2.5' ecoclimatic layers gave the most effective identification of samples resistant to stem rust. By contrast, in experiment 5 (Table 6) the ICARDA-30'' w-pet (with evapotranspiration included) ecoclimatic layers showed the highest PPV, LR+ and estimated Gain when compared to the other ecoclimatic sets. However there was very little difference in predictive performance between the ecoclimatic layers. None of the ecoclimatic layers in experiment 5 produced significant higher performance than any of the other layers (overlapping 95% confidence intervals, LR+, Table 6).

All of the models (ecoclimatic sets) in experiment 5 have substantially higher predictive performance compared to the random selection (by a very good margin not overlapping 95% confidence intervals).

Experiment 6
Since the predictive performance in pre-study 3 for net blotch was significantly lower than that observed for stem rust (compare Tables 1 and 3) we explored the predictive performance for different climate data sets in more detail in this experiment than we did for the stem rust set to see if the performance could be improved.

Table 7: Comparison of the performance of the SIMCA model to select barley accessions resistant to net blotch, when using different ecoclimatic data (and different resolutions) (experiment 6).

Ecoclimate	PPV	LR+	Estimated Gain
ICARDA 30" w-pet (1 km)	0,52 (0,47-0,57)	1,67 (1,43-1,95)	1,32 (1,19-1,45)
ICARDA 30" (1 km)	0,53 (0,47-0,58)	1,69 (1,44-1,99)	1,33 (1,19-1,46)
WorldClim 10' (18.5 km)	0,54 (0,47-0,60)	1,71 (1,37-2,14)	1,33 (1,18-1,50)
WorldClim 5' (9.3 km)	0,53 (0,48-0,59)	1,71 (1,44-2,04)	1,33 (1,20-1,47)
WorldClim 2.5' (4.5 km)	0,56 (0,50-0,61)	1,81 (1,52-2,17)	1,36 (1,23-1,49)
BIOCLIM 10' (18.5 km)	**0,56** (0,48-0,63)	1,89 (1,43-2,53)	1,40 (1,20-1,59)
BIOCLIM 5' (9.3 km)	0,55 (0,46-0,63)	1,86 (1,34-2,56)	1,39 (1,16-1,58)
BIOCLIM 2.5' (4.5 km)	0,55 (0,46-0,64)	**1,90** (1,34-2,72)	1,40 (1,17-1,63)
Random selection	*0,40 (0,35-0,45)*	*0,99 (0,84-1,17)*	*0,99 (0,87-1,12)*

The BIOCLIM ecoclimatic layers are included with the same resolutions. The record level data unit is here the original observation measurement from the USDA net blotch dataset (total 2786 accessions, prevalence = 0.40). The 95% confidence interval is included inside the parentheses.

All the ecoclimatic sets yielded hit-rates, for identification of samples resistant to net blotch, 32% to 40% higher than if selections had been made at random (Table 7). However, while the magnitudes of the LR+ and Estimated Gain for the BIOCLIM data show a marginally better performance than other classes of ecoclimatic data, the confidence intervals indicate there was no appreciable difference between using the different ecoclimatic sets or by using different degrees of resolution for the climatic surfaces (Table 7). That is, the finer resolutions did not improve the predictive performance of the model. Further, the predictive performance for the net blotch dataset still remained significantly poorer than those achieved for the stem rust dataset.

DISCUSSION

This study clearly shows that the ecogeographic distribution of both stem rust resistance in wheat and net blotch in barley is not random, but rather, is linked to climatic factors. This supports the findings of Bonman et al. (2005, 2007) for rust and blotch diseases. Further, if this holds true for stem rust and net blotch then it is reasonable to assume it would hold true for other pests and diseases. As such, we can conclude that variables describing collection site environments can be used to identify disease or pest resistant landraces or wild relative accessions conserved in genetic resource collections at a better frequency than if material is selected at random or with a core collection. Trait mining using FIGS requires a small training set with trait scores to be available for the calibration of the model. While the work of El Bouhssini et al. (2009, 2011) and Bhullar et al. (2009, 2010) showed how this can be done to good effect for both pests and diseases, their studies did not include a comparison to a random selection process. By contrast this study did and thus can be considered as the first definitive proof of concept for the FIGS strategy applied to genebank mining for useful traits.

An important limitation for the exploitation of the link between the ecoclimatic data and the trait property is the requirement of a small set a priori trait data to train the trait mining model. A heuristic approach to incorporate expert knowledge to select samples for the initial training set

will help to reduce this dependence of a priori trait data. However this training set needs to be screened before the first trait-mining model can be calibrated. The further FIGS sets can be developed in a stepwise to incorporate all samples screened each trial season in the trait-mining model to select the samples for the next trial season. Another limitation when using the FIGS strategy is the requirement of georeferenced collecting sites. The ecoclimatic dataset is extracted based on the geographic coordinate. Some of the genebank accessions lack appropriate information required to identify the collecting site.

The indications are that the modeling approaches used in this study could be useful to predict disease resistances in untested germplasm, provided there are data that can be used as a trainer set. However, this study also indicates that the models used here are sensitive to differences in where and when screenings take place (experiment 3 and 4), the pathogen being tested (experiment 1 compared to experiment 5) and the host crop (experiment 2). For the stem rust set the race used as inoculum was also different between trial years and trial season, which likely contribute to the difference in predictive performance observed for experiment 2. The utility of these models in an applied context are still yet to be established. In a follow on study by the same authors the same models and trainer data deployed in this study were used to select small subsets of wheat landrace material with a higher than random frequency of Ug99 resistance. This study included 4563 wheat landraces screened for resistance to stem rust Ug99 in Yemen during 2007. The observed trait scores were not reveled to the person making the trait mining models except from a small training set of 825 samples (20%). The trait-mining model developed using the same method as described here was used to select a subset of 500 samples predicted to have a higher likelihood of resistance for Ug99. The complete set included 10.2% resistant samples while the selected set of 500 samples was found to have 25.8% resistant samples; thus demonstrating that the models and approach used here can indeed be applied to a real life genebank situation (unpublished data, 2010) [PAPER IV]. The challenge now is to improve the robustness and predictive power of the approaches used.

The initial pre-study tests revealed that for these trait datasets (stem rust and net blotch sets) the SIMCA classification algorithm produced the best models. If the datasets had fulfilled the parametric assumptions (normal distribution of variables and residuals) then we would have expected the LDA and PLS-DA algorithms to produce the best models. This is not the case for these datasets. The SIMCA algorithm can be seen as similar to the kNN approach in that it selects the class of PCA model that the sample is most similar to, while kNN selects the class that is most similar to the nearest sample. The pre-study tests revealed that the SIMCA models performed better than the kNN models (Table 1, 2, and 3) and thus would be the model of choice to use in an applied context when the data are not normally distributed.

Reclassification of the measurement scale

For the stem rust set the SIMCA models for different reclassification of the measurement scale indicated that the reclassification to use a scale with three levels improved the predictive performance. However, it must be noted that the reclassification must make sense in terms of what the measurement scale represents. That is, in this case disease scores from 0-3 can be considered as resistant, 4-6 as moderately resistant and 7-9 as susceptible. The original trait measurement scale with 9 or 10 category levels caused problems for the calibration of some of the models (in particular for the stratified subsets with fewer number of samples) because of the lack of samples to represent some of the category levels. Even for the 'training sets' with samples to represent all category levels a high number of levels may cause other issues. One such issue relates to how the classifier relates to the so-called level of measurement. Stevens (1946) suggested a formal taxonomy for different types of measurement scales (nominal, ordinal, interval, and ratio). All of the classifiers we used in this experiment (LDA, PLS-DA,

kNN, and SIMCA) only make use of information from a nominal type measurement scale. This means that the classifiers do not assume any order in the category levels and does not 'know' that the trait score 2 is between trait score 1 and trait score 3. This means that including more category levels in the dataset does not give the classifier more useful information to identify the resistant samples based on the order of the category levels. The reduction to a measurement scale with all the target samples (resistant landraces) grouped together, is likely to give the classifier more information relevant to the task at hand - that is to discriminate the resistant samples from the susceptible samples. Following this argument one might expect that the trait scale with two category levels would be the best alternative. This was not the case in this study (pre-study 1, Table 1; pre-study 2, Table 2) where the S3 scale showed a tendency to give the highest predictive performance (however not statistically significant as evaluated by the overlapping 95% confidence intervals). It is possible that the samples from the original trait scores 7-9 provides the classifier with more coherent examples of the difference between the resistant and the susceptible samples than using original trait scores 4-9 as examples of susceptible samples. A combination of samples from the intermediate resistance and the susceptible groups would thus remove information that the model otherwise was able to exploit; while reducing the original measurement scale to 3 levels only remove information that the model was not able to exploit.

In the final step to evaluate the classification performance the predicted trait scores were reduced to only two levels with the confusion matrix. This last step of the trait mining experiments reported here was however made after the classifier has extracted information from the dataset and was motivated by the primary interest of this study to evaluate the performance of the classifier to identify resistant samples (positives) rather than to distinguish resistant samples from intermediate and susceptible samples.

Stratification by species and screening year and location

For all of the stratification subsets the trait mining models perform better than a random selection. For some subsets the predictive performance is notably higher than for other subsets. But when we compare the predictive performance for different stratified subsets in the same group to the overall average predictive performance, we generally find that some subsets perform better while other subsets, in the same groups of subsets, perform lower than the overall average. For the stem rust set (Table 5) we found that some of the models limited to one single species, experiment site, or trial year, perform better than the overall average performance, while others in the same group perform lower than the overall average (SIMCA, S3, Table 1). Similar tests in the net blotch set (not reported, see supplementary material, S1.5) with stratification by experiment site and year, show the same pattern. Thus we did not observe any clear evidence that splitting the full dataset into smaller stratified subsets, containing similar samples, provides any general improvement in predictive performance. In terms of creating data subsets across different models, there is still the question: does adding more predictive variables (more ecogeographical layers, see below) and splitting the data to different genetic backgrounds improve the prediction?

The BIOCLIM ecoclimatic layers

The BIOCLIM variables (Busby, 1991) are derived from the raw climate variables with the aim to better describe the ecoclimatic environment with parameters that are more directly relevant for the description of the ecological niche. However, for the trait mining experiments with the stem rust set, the predictive performance is significantly higher when using the raw climate variables rather than the derived BIOCLIM variables (not overlapping 95% CI for the WorldClim layers compared to the BIOCLIM layers in experiment 1, Table 4). Note that in the net blotch set this result is reversed, and here the BIOCLIM ecoclimatic set calibrates the

models with the highest predictive performance (experiment 6, Table 7). The original purpose of developing the BIOCLIM variables was for the calibration of envelope models where the maximum and minimum value for each of the ecoclimatic variables is assumed to define the boundaries of the species habitat (Busby, 1991; Franklin, 2010). For this study we have not explored the envelope modeling principle, but rather used standard multivariate classification methods. Our study indicates that for the stem rust set the raw ecoclimatic variables perform better than the derived BIOCLIM variables when using the SIMCA classifier. It is possible that the processing of the raw climate variables into the BIOCLIM variables does not always preserve all the predictive information content for our approach.

The potential value of additional ecogeographic layers

The second observation related to the ecoclimatic variables is that when the potential evapo-transpiration (PET), from the ICARDA ecoclimatic database, is added to the ecoclimatic variables, the predictive performance is sometimes slightly higher (Table 6). This climate variable is only available for the ICARDA ecoclimatic dataset. When we repeat the same trait mining experiments with potential evapo-transpiration included or excluded, we find for experiment 5 (Table 6) (but not for experiment 1 (Table 4) and experiment 6 (Table 7)) that including this ecoclimatic property improves the predictive performance. This indicates that the PET climatic layer could in some contexts carry independent predictive information that is useful to the trait mining models. This is hardly surprising in that atmospheric humidity around the plant directly impacts the PET and it is widely accepted that high humidity is associated with infection by fungal type plant pathogens (eg Hoffmann and Schmutterer, 1983).

Whatever the underlying reason, this result leads us to suggest that choosing appropriate ecoclimatic variables will be crucial to improving the predictive performance of FIGS. For example, in this study monthly variables were used that described the entire year from January to December. However, for a given collection site the growing season usually does not start in January and does not last the whole year thus making some monthly variables not as relevant as others. A suggestion for further FIGS studies of this kind would thus be to explore the effects on the predicative power by 1) aligning monthly variable according to onset of growing period 2) only including monthly variables that are within the growing period for a given site. This would however necessitate accurate estimations of the onset of growing period. Continuous surfaces for this variable have been developed at ICARDA by De-Pauw et al. (personal communication) and are currently being used in a study as suggested here.

Predictive performance is lower for net blotch than for stem rust

Although we see a similar pattern of the performance indicators when compared to the random selection, the predictive performance indicators are substantially lower for the net blotch set than they are for the stem rust set. Afanasenko et al. (1995) discovered that the resistance in barley against net blotch caused by *Pyrenophora teres* f. *teres* (net form) and *Pyrenophora teres* f. *maculata* (spot form) are inherited independently. Perhaps resistance to net blotch in barley is more complex than the resistance to stem rust in wheat and thus more difficult to capture using the models developed in this study. Bonman et al. (2005) found that the response to net blotch in this dataset is correlated to the winter habit of the germplasm samples. It is possible that a trait mining study on subsets for each winter habit separate would give a higher predictive performance. However, overall both datasets show a very satisfactory predictive performance for the FIGS strategy in this study.

Distinct collecting sites

Many genebanks contain multiple accessions from the same collection site. This could be due to a variety of reasons including multiple accessions being collected from the same site, one

accession being split into different genotypes, non-geo-referenced accessions from a given province being assigned a collection site geo-coordinate that corresponds with the central point of the province or material being stored in a collection are assigned collection site geo-coordinates corresponding to the physical location of the collection. Clearly, to use accessions where the latter example is the case in a FIGS analysis would not be appropriate. However the FIGS approach is relevant if the collection site geo-coordinates are reasonably accurate thus the question then becomes how one treats multiple accessions per collection site in when using a FIGS approach.

The results of experiment 5 (Table 6) suggest that, for analysis such as those reported, it is better to use an average score across accessions from the same site so that the site representation in the analysis is kept to a single entry. Many of the collecting sites in the stem rust set have a very high number of accessions collected at the same site. During the calibration of the classification models these collecting sites provide a very high number of examples from which the models could learn. The models could thus be biased, focusing too much on these collecting sites and neglecting useful information from the collecting sites with fewer accessions. Another contributing explanation could perhaps be that some of the accessions from the same collecting site have very different trait score values. The calibration of classification models would thus receive a number of conflicting examples linking the same ecogeographic pattern to both high trait scores and to low trait scores. For the dataset with distinct sites, the average of the different trait scores for each site gave the calibration routine only one example of the link between the ecogeography and trait score for each site.

Different resolutions for the ecoclimatic layers

A somewhat unexpected result was that the finer spatial resolution for the ecoclimatic layers did not improve the predictive performance (experiment 6, Table 7). This experiment was only made for the net blotch set. Even if the predictive performance was slightly higher for the finer resolutions, the 95% confidence intervals clearly show that the observed differences between the different resolutions are insignificant. In mountainous areas the differences in particular for temperature, but also for precipitation can be substantial within the different spatial resolutions explored. However the area of (adaptive) cultivation for a landrace will sometimes be larger than even the largest grid cell we explored (10 arc-minutes = 18.5 km = 342 km^2). It is further possible that the georeferenced coordinates for the reported collecting site is not the center-point for the area of cultivation for the landraces. In some cases the collecting site might even be a farmers market in close proximity to the typical cultivation area. For these examples the ecoclimate of the coarser spatial resolutions might be a better representative of the typical ecoclimate of the landrace than the ecoclimate of the smaller resolution pixel centered at the collecting site. Further experiments to explore the effect of the spatial resolution for the ecoclimatic layers would be useful.

Assumptions for the FIGS approach

The FIGS strategy is based on the assumption that the expression of a useful trait, for example pest resistance, in landraces (and crop wild relatives) is linked to the environmental parameters describing the collection site, and that we can build a statistical model to define a signature for the ecogeography of these landraces. The model, in this study, is applied as a search pattern to identify other landraces originating from locations with similar ecogeography as the resistant landraces. In practice these landraces would be selected as candidate samples for a field trial to screen for the target trait. Trait mining with FIGS aims to identify a higher proportion of resistant landraces than would be expected without the application of this selection strategy (Mackay and Street, 2004).

When modeling the crop resistance against a pathogen it is important to remember that the distribution of the pathogen is directly linked to the ecogeography. For example, given many pathogens are sensitive to humidity; it is possible that the improved performance of the SIMCA model demonstrated in experiment 5 was due to the inclusion of the evapo-transpiration parameters. For pathogens like stem rust the distribution of the alternative host, barberry (Berberis L.), is required for sexual reproduction of the pathogen. The virulence of the pathogen is thus expected to be higher in areas where barberry grows in the proximity of the cultivated crop plants. The predictive association between the resistance trait and the ecoclimatic variables we have identified with this study is thus, at least partly, an indirect link. The models are likely to describe the suitable ecogeography where the pathogen thrives and thus are likely to impose a selection pressure for the emergence of resistance genes within *in situ* populations. This was illustrated by Paillard et al. (2000) who report that populations of winter wheat with the highest level of resistance to powdery mildew originated from sites where powdery mildew pressure was high, due to environmental factors, while the reverse was true of those populations where the pressure was low. On the other hand, the models reported here are less likely to describe the ecoclimatic conditions favorable for the crop to develop traits that would protect it against the pathogen, without the presence of the pathogen. The development of useful resistance in the landraces is an adaptive response to the biotic stress from the pathogen and not the environment per se. However to complicate the picture further, Stukenbrock and McDonald (2008) pointed out that many crop pathogens have been domesticated together with their host crop and are thus linked back to the geographic distribution of the crop.

The most important aspect of FIGS is that it is predictive. FIGS does not aim to describe the mechanism behind the crop traits. If the models used to develop FIGS sets are predictive then they could be used to develop smaller subsets with a higher hit-rate for a targeted crop trait.

FIGS models provide a complement to expert knowledge

The FIGS approach is not intended to replace the valuable expert knowledge held by crop breeders and genebank curators. When planning a new field experiment, the predictions from FIGS will assist the crop expert to select the most appropriate genebank accessions to include. The size of the smaller subset could be limited by the capacity given by the size of the available field area, laboratory capacity, or by the project funding available for human resources.

Possible causes of (eventual) prediction problems

The predictive performance for the experiments in the stem rust and net blotch set from this current study was good. However if the predictive performance is low when the approach described here is followed, the list below provides some suggestions on how to improve the hit-rate.

- The algorithm of the classifier is not able to recognize and discriminate all the samples. Further additional classification methods and other pre-processing methods can be explored.
- The models explored and compared in this study used long-term monthly climatic data arranged from January to December. However, when refining these processes it will be interesting to test if the predictive power of the techniques is improved when start of growing seasons are aligned so that only those months in which the crop would normally develop *in situ* are used in the models. In other words sites are agro-climatically compared for similarities based on conditions prevailing during the actual growing seasons.
- The ecogeographic data from the source location of the landrace does not contain enough relevant information that could be linked to the evolution of a given crop trait. Other ecogeographic datasets or other grid resolutions can be explored. For example,

measures of long-term season-to-season variation for climatic parameters would be useful when considering crop adaptation strategies.

- Data quality, precision or error issues of the germplasm passport data. Data quality is of course paramount in any data analysis study. Written logbooks, collecting mission reports, and similar sources can be revisited to complete missing data and improve on data accuracy, particularly the precision of collection site geo-coordinates.
- Lack of replicated measurements. Many datasets with evaluation of genebank material includes only one single observation for each genebank accession. With the lack of replication across multiple experiment years and experiment locations it is very difficult to assess the precision and to estimate the natural variance of the trait scores, or observations. It is also difficult to estimate any GxE (genotype by environment interaction) effects in the trait dataset. Care should thus be made whenever possible to include replicated measurements across both experiment site and year, for future trait evaluations. It is also important to apply an appropriate sampling design to avoid systematic bias in the recorded data.
- Assessment of trait variation could also be a problem as trait observations might include unexpected bias and mistakes. Some of the individual observations from the crop trait training-set could be the result of an unusual experimental condition. As for example locally higher pest stress pressure in smaller parts of the field plot, or unusual low or high pest activity during some of the trial seasons. With the absence of repetitions it is difficult to evaluate this aspect. The initial data analysis can be made to explore outliers and to identify the most important problem samples. However outliers can be valid data points and should not always be removed.
- When working with cultivated material (like landraces), the adaptive development of the crop trait might be more dominantly explained by the breeding decisions made by the farmer. For more modern cultivated material there is no appropriate location of origin, as the breeding lines are often the complex result of crossing between genetic resources from very many different source locations. For this problem FIGS strategy may not be the most appropriate approach.

Future work

The predictive performance from other different classification methods should be explored. In this study we found significant variation between the 4 different classification methods we used. The Artificial Neural Networks (ANN; Bishop, 1996) is one particular interesting method to explore because the algorithm is so different from the algorithms of the methods used here - and also because the failure of the linear discriminate analysis (LDA; Fisher, 1936) classifier indicates that the classification problem here is not typical for a parametric solution. Another classifier that could be explored is Decision Tree methods like the Random Forest algorithm (Breiman, 2001; Stockwell, 2007). With multiple different classification methods the so-called ensemble classifier method (Kuncheva, 2004; Rokach, 2010) could be used to combine the predictive information from each classifier. The classifier ensemble will often provide a higher predictive performance than even the best individual classifier. The performance of the classifier ensemble is based on the assumption that each classifier describes the dataset independently and in a different way from the other classifiers.

A significant amount of work still needs to be done to identify or create environmental parameters that are more tightly linked to the evolution of traits so that the predictive power of models can be improved.

Another obvious use case would be to apply the FIGS approach to analyzing gaps in genebank collections. The ecogeographic signature for a particular crop trait can be applied to identify

likely locations with specific genetic diversity not yet represented in the collection. FIGS could thus guide new collecting expeditions to interesting new locations based particular target crop traits. This can be compared to similar gap analysis studies with species distribution models, where the purpose is to complete the genebank collection with overall genetic diversity not yet represented in the collection (Jarvis et al., 2003; and Jarvis et al., 2005, Upadhyaya et al., 2009; Ramírez-Villegas et al., 2010). This use case for the FIGS strategy can be seen as a natural extension of the ecological niche modeling methods to estimate species' distributions.

CONCLUSION

This study contributes to the development of methods for identifying FIGS subsets of georeferenced genebank accessions more likely to contain sought after novel genetic variation for adaptive traits. The objective of the FIGS strategy is to more efficiently identify and utilize plant genetic resources, particularly the landrace and wild relatives of crop plants. The results support the assertion that trait mining using the FIGS approach can significantly improve the hit-rate for identification of landrace samples with resistance to target crop pests. FIGS subset selection is proposed as an alternative approach to the selection of a core collection to assess rare and useful traits such as resistance to diseases, pests and abiotic constraints.

Acknowledgments

Dr Harold E. Bockelman head of the USDA ARS National Small Grain Collection in Aberdeen, Idaho extracted the stem rust and net blotch dataset from the USDA NPGS GRIN database. Associate professor Dvora-Laiô Wulfsohn (Copenhagen University) provided feedback and suggestions to the draft manuscript. Dr Axel Diederichsen and other colleagues at NordGen provided challenging feedback and discussions on the research topics of this manuscript. Many thanks also to the late Dr Bent Skovmand (NordGen) for help and advice when this research project was started. This research project is supported by a grant from the Nordic Genetic Resources Center (NordGen, www.nordgen.org).

Author contributions

All authors helped to assemble the data and to develop the experimental design for the modeling studies. DE made the data analysis and wrote the first version of the manuscript. All authors contributed to the final manuscript.

REFERENCES

Altman, D.G., and J.M. Bland (1994a). Statistical Notes: Diagnostic tests 1: sensitivity and specificity. BMJ 308(6943): 1552.

Altman, D.G., and J.M. Bland (1994b). Statistical Notes: Diagnostic tests 2: predictive values. BMJ 309(6947): 102.

Agresti, A. (2002). Categorical Data Analysis. Second edition. John Wiley and Sons, Hoboken, New Jersey, USA. ISBN: 9780471360933.

Afanasenko, O.S., H. Hartleb, N.N. Guseva, V. Minarikova, M. Janosheva (1995). A Set of Differentials to Characterize Populations of *Pyrenophora teres* Drechs. for International Use. Journal of Phytopathology 143(8): 501-507. DOI: 10.1111/j.1439-0434.1995.tb04562.x.

Barker, M., and W. Rayens. 2003. Partial least squares for classification. J. Chemometrics 17(3): 166-173. DOI: 10.1002/cem.785.

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: 9519-9524. DOI: 10.1073/pnas.0904152106.

Bhullar, N.K., Z. Zhang, T. Wicker, and B. Keller (2010). Wheat gene bank accessions as a source of new alleles of the powdery mildew resistance gene *Pm3*: a large scale allele mining project. BMC Plant Biology 10: 88. DOI: 10.1186/1471-2229-10-88.

Bishop, C. (1996). Neural Networks for Pattern Recognition. Oxford University Press, UK. ISBN: 978-0198538646.

Bonman, J.M., H.E. Bockelman, L.F. Jackson, and B.J. Steffenson (2005). Disease and Insect Resistance in Cultivated Barley Accessions from the USDA National Small Grains Collection. Crop Science 45:1271-1280. DOI: 10.2135/cropsci2004.0546.

Bonman, J.M., H.E. Bockelman, Y. Jin, R.J. Hijmans, and A.I.N. Gironella (2007). Geographic Distribution of Stem Rust Resistance in Wheat Landraces. Crop Science 47: 1955-1963. DOI: 10.2135/cropsci2007.01.0028.

Borlaug Global Rust Initiative. Available at http://www.globalrust.org (verified 21 Dec 2010).

Breiman, L. (2001). Random Forests. Machine Learning 45(1): 5–32. DOI: 10.1023/A:1010933404324.

Brown, A.H.D., and C. Spillane (1999). Implementing core collections principles, procedures, progress, problems and promise. p. 1-9. *In:* Johnson, R.C., and T. Hodgkin (ed). Core collections for today and tomorrow. International Plant Genetic Resources Institute, Rome.

Busby, J.R. (1991). BIOCLIM - A Bioclimatic Analysis and Prediction System. p. 64-68. *In*: Margules, C.R., and M.P. Austin (eds). Nature Conservation: Cost Effective Biological Surveys and Data Analysis. CSIRO, Canberra, Australia.

Cardillo, G. (2007). Cohen's kappa: Compute the Cohen's kappa ratio. Available online at http://www.mathworks.com/matlabcentral/fileexchange/15365 (MATLAB script, downloaded on 27 July 2010, verified 21 Dec 2010).

CEBM (2010). Statistics calculator [Online]. Center for Evidence-Based Medicine, University Health Network. Available online at http://ktclearinghouse.ca/cebm/practise/ca/calculators/statscalc (verified 21 Dec 2010).

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement 20: 37-46. DOI: 10.1177/001316446002000104.

Cover, T.M., and P.E. Hart (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory 13(1): 21-27. DOI: 10.1109/TIT.1967.1053964.

De Pauw, E. (2008). Climatic and Soil Datasets for the ICARDA Wheat Genetic Resource Collections of the Eurasia Region. Explanatory Notes. ICARDA GIS Unit, Aleppo, Syria. 68 p. Available at http://geonet.icarda.cgiar.org/geonetwork/data/regional/GRU_NetBlotch/Doc/Report_NetBlotch.pdf (6.6 MB, verified 21 Dec 2010).

Dorofeyev, V.F. (ed) (1992). N. I. Vavilov, Origin and Geography of Cultivated Plants. Cambridge University Press, Cambridge, UK. ISBN: 978-0-521-11159-1. [Translated from Russian by D. Löve]

Duda, R.O., P.E. Hart, and D.G. Stork (2001). Pattern classification 2nd edition. Wiley, University of Michigan, USA. ISBN: 9780471056690.

Dwivedi, S.L., J.H. Crouch, D.J. Mackill, Y. Xu, M.W. Blair, M. Ragot, H.D. Upadhyaya, and R. Ortiz (2007). The molecularization of public sector crop breeding: Progress, problems, and prospects. Advances in Agronomy 95: 163-318.

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

El Bouhssini, M., K. Street, A. Amri, M. Mackay, F.C. Ogbonnaya, A. Omran, O. Abdalla, M. Baum, A. Dabbous, and F. Rihawi (2011). Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). Plant Breeding 130(1): 96-97. DOI: 10.1111/j.1439-0523.2010.01814.x.

Endresen, D.T.F. (2010). Predictive association between trait data and ecogeographic data for Nordic barley landraces. Crop Science 50(6): 2418-2430. DOI: 10.2135/cropsci2010.03.0174.

FAO (1974). FAO-UNESCO Soil Map of the World. Vol. I: Legend. UNESCO, Paris.

FAO (2002). International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA). FAO, Rome, Italy.

FAO (2007). Digital Soil Map of the World [computer file]. Version 3.6. Available at http://www.fao.org/geonetwork/srv/en/metadata.show?id=14116 (verified 21 Dec 2010).

FAO (2010). The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. FAO, Rome, Italy. ISBN: 978-92-5-106534-1.

FAO-UNESCO (1995). The Digital Soil Map of the World and Derived Soil Properties [CD-ROM computer file]. Land and Water Digital Media Series 1. FAO, Rome. Information available at http://www.fao.org/ag/agl/lwdms.stm#cd1 (verified 21 Dec 2010).

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Annals of Eugenics 7(2): 179-188.

Flood, J. (2010). The importance of plant health to food security. Food Security 2(3): 215-231. DOI: 10.1007/s12571-010-0072-5.

Frankel, O. (1984). Genetic perspectives of germplasm conservation. p. 161-170. *In:* Arber, W., K. Illmensee, W.J. Peacock, and P. Starlinger (eds). Genetic manipulation: Impact on man and society. Published on behalf of the ICSU Press by Cambridge University Press, Cambridge, UK. ISBN: 0521264170.

Franklin, J. (2010). Mapping Species Distributions. Spatial Inference and Prediction. Cambridge University Press, Cambridge, UK. ISBN: 978-0-521-70002-3.

Gepts, P. (2006). Plant genetic resources conservation and utilization: The accomplishments and future of a societal insurance policy. Crop Science 46: 2278-2292. DOI: 10.2135/cropsci2006.03.0169gas.

Gwet, K.L. (2010). Handbook of Inter-Rater Reliability (Second Edition), The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters. Advanced Analytics, LLC, Gaithersburg, MD, USA. 208 pages. ISBN: 9780970806222.

Hawkins, D.M. (2004). The Problem of Overfitting. Journal of Chemical Information and Computer Sciences 44(1): 1-12. DOI: 10.1021/ci0342472.

Hawkins, D.M., S.C. Basak, and D. Mills (2003). Assessing Model Fit by Cross-Validation. Journal of Chemical Information and Computer Sciences 43(2): 579–586. DOI: 10.1021/ci025626i.

Hijmans, R.J., S.E. Cameron, J.L. Parra, P.G. Jones, and A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology 25: 1965-1978. DOI: 10.1002/joc.1276.

Hijmans, R.J., M. Jacobs, J.B. Bamberg and D.M. Spooner (2003). Frost tolerance in wild potato species: Unraveling the predictivity of taxonomic, geographic, and ecological factors. Euphytica 130: 47–59. DOI: 10.1023/A:1022344327669.

Hintum, Th.J.L. van, A.H.D. Brown, C. Spillane, and T. Hodgkin (2000). Core collections of plant genetic resources. IPGRI Technical Bulletin No. 3. International Plant Genetic Resources Institute, Rome, Italy. ISBN: 9789290434542.

Hoffman, G.M. and H. Schmutterer. 1983. Parasitäre Krankheiten und Schädlinge an landwirtschaftlichen Kulturpflanzen. Eugen Ulmer GmbH & Co., Stuttgart, Germany. ISBN: 3-8001-3058-0.

Holbrook, C.C., and W. Dong (2005). Development and Evaluation of a Mini Core Collection for the U.S. Peanut Germplasm Collection. Crop Science 45:1540-1544.

Hutchinson, M.F. (1995). Interpolating mean rainfall using thin plate smoothing splines. International Journal of Geographical Information Systems 9: 385-403.

Hutchinson, M.F. (2000). ANUSPLIN version 4.1. User Guide. Center for Resource and Environmental Studies, Australian National University, Canberra, Australia.

Jalli, M. (2010a). The virulence of Finnish *Pyrenophora teres* f. *teres* isolates and its implications for resistance breeding. Ph.D diss. MTT Agrifood Research Finland. ISBN: 978-952-487-274-4. Available at http://www.mtt.fi/mtttiede/pdf/mtttiede9.pdf (2.2 MB, verified 21 Dec 2010).

Jalli, M. (2010b). Sexual reproduction and soil tillage effects on virulence of *Pyrenophora teres* in Finland. Annals of Applied Biology (online in advance of print). DOI: 10.1111/j.1744-7348.2010.00445.x.

Jarvis, A., M.E. Ferguson, D.E. Williams, L. Guarino, P.G. Jones, H.T. Stalker, J.F.M. Vallis, R.N. Pittman, C.E. Simpson, and P. Bramel (2003). Biogeography of wild *Arachis*: Assessing conservation status and setting future priorities. Crop Science 43(3): 1100-1108. DOI: 10.2135/cropsci2003.1100.

Jarvis, A., K. Williams, D. Williams, L. Guarino, P.J. Caballero, and G. Mottram (2005). Use of GIS for optimizing a collecting mission for a rare wild pepper (*Capsicum flexuosum* Sendtn.) in Paraguay. Genetic Resources and Crop Evolution 52: 671-682. DOI: 10.1007/s10722-003-6020-x.

Koerner, B.I. (2010). Red Menace: Stop the Ug99 Fungus Before Its Spores Bring Starvation. WIRED Magazine, March 2010. Available at http://www.wired.com/magazine/2010/02/ff_ug99_fungus/all/1 (verified 21 Dec 2010).

Kohavi, R., and F. Provost (1998). Glossary of terms. Machine Learning 30(2-3): 271-274. DOI: 10.1023/A:1017181826899.

Krupinsky, J.M., K.L. Bailey, M.P. McMullen, B.D. Gossen, and T.K. Turkington (2002). Managing Plant Disease Risk in Diversified Cropping Systems. Agronomy Journal 94(2): 198–209. DOI: 10.2134/agronj2002.1980.

Kuncheva, L.I. (2004). Combining Pattern Classifiers. Methods and Algorithms. John Wiley & Sons, Hoboken, New Jersey, USA. ISBN: 0-471-21078-1.

Kurlovich, B.S., S.I. Rep'ev, M-V. Petrova, T.V. Buravtseva, L.T. Kartuzova, and T.A. Voluzneva. (2000). The significance of Vavilov's scientific expeditions and ideas for development and use of legume genetic resources. Plant Genetic Newsletter 124: 23-32.

Landis, J.R., and G.G. Koch (1977). The measurement of Observer Agreement for Categorical Data. Biometrics 33(1): 159–174. Stable URL: http://www.jstor.org/stable/2529310 (verified 21 Dec 2010).

Liu, Z., S.R. Ellwood, R.P. Oliver, and T.L. Friesen (2011). *Pyrenophora teres*: profile of an increasingly damaging barley pathogen: Pathogen profile of *Pyrenophora teres*. Molecular Plant Pathology 12(1): 1-19. DOI: 10.1111/j.1364-3703.2010.00649.x.

Liu, Z.H., and T.L. Friesen (2010). Identification of *Pyrenophora teres* f. *maculata*, Causal Agent of Spot Type Net Blotch of Barley in North Dakota. Plant Disease 94(4): 480-480. DOI: 10.1094/PDIS-94-4-0480A.

Mackay, M.C. (1990). Strategic planning for effective evaluation of plant germplasm. p. 21–25. *In*: Srivastava J.P., and A.B. Damania (eds). Wheat Genetic Resources: Meeting Diverse Needs. John Wiley & Sons, Chichester, UK. ISBN 0-471-92880-1.

Mackay, M.C. (1995). One core collection or many? p. 199-210. *In:* Hodgkin T., A.H.D. Brown, Th.J.L. van Hintum, and A.A.V. Morales (eds). Core Collections of Plant Genetic Resources. John Wiley & Sons, Chichester, UK. ISBN: 471-95545-0.

Mackay M.C., and K. Street (2004). Focused identification of germplasm strategy – FIGS. p. 138-141. *In*: Black, C.K., J.F. Panozzo, and G.J. Rebetzke (eds). Proceedings of the 54th Australian Cereal Chemistry Conference and the 11th Wheat Breeders' Assembly. Royal Australian Chemical Institute, Melbourne, Australia.

McIntosh, R.A., C.R. Wellings, and R.F. Park (1995). Wheat Rusts: An Atlas of Resistance Genes. CSIRO, Melbourne, Victoria, Australia. ISBN: 0-643-05428-6.

McLean, M.S., B.J. Howlett, and G.J. Hollaway (2009). Epidemiology and control of spot form of net blotch (*Pyrenophora teres* f. *maculata*) of barley: a review. Crop Pasture Sci. 60(4): 303-315. DOI: 10.1071/CP08173.

Negri, V, N. Maxted, and M. Veteläinen. 2009. European landrace conservation: an introduction. p. 1-22. In: Veteläinen, M., V. Negri, and N. Maxted. 2009. European landraces on-farm conservation,

management and use. Bioversity Technical Bulletin No. 15. Bioversity International, Rome, Italy. ISBN: 978-92-9043-805-2.

Paillard S., I. Goldringer, J. Enjalbert, M. Trottet, J. David, C. de Vallavieille-Pope, and P. Brabant (2000) Evolution of resistance against powdery mildew in winter wheat populations conducted under dynamic management- II: Adult plant resistance. Theor Appl Genet 101: 457-462. DOI: 10.1007/s001220051503.

Pessoa-Filho, M., P.H.N. Rangel, and M.E. Ferreira (2010). Extracting samples of high diversity from thematic collections of large gene banks using a genetic-distance based approach. BMC Plant Biology 10:127.

Peeters, J.P. and J.T. Williams (1984). Towards better use of genebanks with special reference to information. Plant Genetic Resources Newsletter 60: 22–32. DOI: 10.1007/BF00224023.

Pretorius, Z.A., R.P. Singh, W.W. Wagoire, and T.S. Payne (2000). Detection of virulence to wheat stem rust resistance gene Sr31 in *Puccinia graminis* f. sp. *tritici* in Uganda. Plant Disease 84(2): 203. DOI: 10.1094/PDIS.2000.84.2.203B.

Ramírez-Villegas J., C. Khoury, A. Jarvis, D.G. Debouck, and L. Guarino (2010). A Gap Analysis Methodology for Collecting Crop Genepools: A Case Study with *Phaseolus* Beans. PLoS ONE 5(10): e13497. DOI: 10.1371/journal.pone.0013497.

Rokach, L. (2010). Pattern Classification using Ensemble Methods. Series in Machine Perception and Artificial Intelligence - Vol. 75. World Scientific Publishing Co. Pte. Ltd., Singapore. ISBN: 9814271063.

Shapiro, S.S., and M.B. Wilk (1965). An analysis of variance test for normality (complete samples). Biometrika 52(3-4): 591–611. DOI: 10.1093/biomet/52.3-4.591.

Silvar, C., A.M. Casas, D. Kopahnke, A. Habekuß, G. Schweizer, M.P. Gracia, J.M. Lasa, F.J. Ciudad, J.L. Molina-Cano, E. Igartua, and F. Ordon (2010). Screening the Spanish Barley Core Collection for disease resistance. Plant Breeding 129(1): 45-52. DOI: 10.1111/j.1439-0523.2009.01700.x.

Singh, P.R., D.P. Hodson, J. Huerta-Espino, Y. Jin, P. Njau, R. Wanyera, S.A. Herrera-Foessel, and R.W. Ward (2008). Will Stem Rust Destroy the World's Wheat Crop? Advances in Agronomy 98: 271-309. DOI: 10.1016/S0065-2113(08)00205-8.

Stakman, E.C. (1915). Relation between Puccinia graminis and plants highly resistant to its attack. Journal of Agricultural Research 4: 193-99.

Stakman, E.C., and F.J. Piemeisel (1917). Problems in preventing plant disease epidemics. American Journal of Botany 44(3): 259-267. Stable URL: http://www.jstor.org/stable/2438808 (verified 21 Dec 2010).

Steffenson, B.J. (1997). Net blotch. p. 28–31. *In:* Mathre, D.E. (ed.) Compendium of barley diseases The American Phytopathological Society, St Paul, MN, USA.

Stevens, S.S. (1946). On the Theory of Scales of Measurement. Science 103(2684): 677-680. DOI: 10.1126/science.103.2684.677.

Stockwell, D. (2007). Niche Modeling: Predictions from Statistical Distributions. Chapman and Hall/CRC. ISBN: 9781584884941.

Stukenbrock, E.H., and B.A. McDonald (2008). The Origins of Plant Pathogens in Agro-Ecosystems. Annual Review of Phytopathology 46: 75-100.

Upadhyaya, H.D., K.N. Reddy, M. Irshad Ahmed, and C.L.L. Gowda (2009). Identification of geographical gaps in the perl millet germplasm conserved at ICRISAT genebank from West and Central Africa. Plant Genetic Resources: Characterization and Utilization 8(1): 45-51. DOI: 10.1017/S147926210999013X.

USDA NPGS GRIN. United States Department of Agriculture, National Plant Germplasm System, Germplasm Resources Information Network. Available at http://www.ars-grin.gov/npgs/ (verified 21 Dec 2010).

Vavilov, N.I. (1920). Zakon gomologicheskikh ryadov v nasledstvennoj izmenchivosti. [The law of homologous series in variation]. Proceedings of the III All-Russian plant breeding conference. Saratov, 16 p. [In Russian] [Cited in Loskutov, 1999:82]

Vavilov, N.I. (1922). The law of homologous series in variation. Journal of Genetics, 12(1): 47-89. DOI: 10.1007/BF02983073.

Vavilov, N.I. (1932). Problema Novykh Kul'tur [Problems concerning new crops]. Sel'chozgiz, Moscow-Leningrad, USSR). p. 256-285. *In*: Dorofeyev, V.F. (ed) (1992). Origin and Geography of Cultivated Plants, Cambridge University Press, Cambridge, UK. ISBN: 978-0-521-11159-1. [Translated from Russian by D. Löve]

Vavilov, N.I. (1935). The phyto-geographical basis for plant breeding. Studies of the original material used for plant breeding. (First published in 1935 in Teoreticheskie osnovy selektsii [Theoretical basis of plant breeding]. Moscow-Leningrad, USSR). p. 316-366. *In*: Dorofeyev, V.F. (ed) (1992). Origin and Geography of Cultivated Plants, Cambridge University Press, Cambridge, UK. ISBN: 978-0-521-11159-1. (Translated from Russian by D. Löve).

Vavilov, N.I. (1957). Mirovye resurcy sortov chlebnych zlakov, zernovych bobovych, l'na i ich ispol'zovanie v selekcii. Opyt agroklimatičeskogo obozrenija važnejŝich polevych kultur. [World resources of cereals, grain leguminous crops and flax and their utilization in plant breeding. Agroecological survey of the principal field crops]. Izdatel'stvo Akademii Nauk SSR, Moskva, Leningrad, 463 p. (In Russian).

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. p. 391-420. *In*: Krishnaiaah, P.R. (ed). Multivariate Analysis. New York: Academic Press, USA.

Wold, S. (1976). Pattern recognition by means of disjoint principal component models. Patt. Recog. 8: 127-139.

Wold, S., and M. Sjostrom (1977). SIMCA: A method for analyzing chemical data in terms of similarity and analogy. p. 243-282. *In*: Kowalski, B.R. (ed). Chemometrics Theory and Application, American Chemical Society Symposium Series 52. American Chemical Society, Washington D.C., USA.

Wold, S., A. Ruhe, H. Wold, and W.J. Dunn (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comp. 5: 735-743.

Wright, S., and B. Gaut (2005). Molecular Population Genetics and the Search for Adaptive Evolution in Plants. Molecular Biology and Evolution 22(3): 506-519.

WTO (1994). Agreement on Trade-Related Aspects of Intellectual Property Rights (TRIPS). Annex 1C of the Marrakesh Agreement Establishing the World Trade Organization (WTO), signed in Marrakesh, Morocco on 15 April 1994. Available at http://www.wto.org/english/docs_e/legal_e/27-trips_01_e.htm (verified 21 Dec 2010).

Xu, Y. (2010). Plant Genetic Resources: Management, Evaluation and Enhancement. p. 151-194. *In*: Molecular Plant Breeding. CABI. ISBN 978-1-84593-392-0.

Supplementary Material

The following supplementary material was submitted together with the manuscript to Crop Science to be available for download from the Journal web site.

Further detailed performance indicators
Additional performance indicators not reported in the manuscript includes Cohen's Kappa, proportion observed agreement (PO), proportion observed positive agreement (PA), proportion observed negative agreement (NA), sensitivity, specificity, Yule's Q, negative predictive value (NPV), negative diagnostic likelihood ratio (LR-).

- S1_tables.pdf (299 KB, 5 pages)

The stem rust dataset
S2_stem_rust.zip (5.9 MB, 3 files)

- S2_descriptors_sr.pdf (111 KB, 4 pages, descriptor list, stem rust set)
- S2_stem_rust_set.txt (4.5 MB, stem rust data, including passport, trait, and ecoclimate)
- S2_stem_rust_set.xlsx (5.1 MB, stem rust set as MS Excel spreadsheet)

The net blotch dataset
S3_net_blotch.zip (4.6 MB, 3 files)

- S3_descriptors_nb.pdf (123 KB, 5 pages, descriptor list, net blotch set)
- S3_net_blotch_set.txt (2.9 MB, net blotch data, incl. passport, trait, and ecoclimate)
- S3_net_blotch_set.xlsx (4.0 MB, net blotch set as MS Excel spreadsheet)

Cover photo submission:

Wheat spikes from plants grown at Alnarp. The reported data analysis is a desktop study and could be illustrated with these wheat spikes on my desk.

Photo by Dag Endresen, 20 June 2004

This photo is similar as the photo of barley spikes chosen as cover photo for Crop Science volume 50, no 6, and submitted together with my previous manuscript on the Focused Identification of Germplasm Strategy (FIGS) published in this issue of Crop Science.

See also:
http://www.flickr.com/photos/dag_endresen/4261640621/

Paper IV

Endresen, Dag Terje Filip, Kenneth Street, Michael Mackay, Abdallah Bari, and Eddy De Pauw (*draft manuscript*). Sources of resistance in bread wheat to stem rust (Ug99) identified using the Focused Identification of Germplasm Strategy (FIGS).

This draft manuscript constitutes my contribution to a collaborative paper with a follow-up to PAPER III by the same authors.

Sources of resistance to stem rust (Ug99) in bread wheat and durum wheat identified using Focused Identification of Germplasm Strategy (FIGS).

DRAFT MANUSCRIPT

Dag Terje Filip Endresen*, Kenneth Street, Michael Mackay, Abdallah Bari, and Eddy De Pauw

This draft manuscript constitutes my contribution to a planned collaborative paper as a follow-up to PAPER III included in this thesis. The collaborative paper will be prepared soon and will include my contributions based on this draft manuscript.

ABSTRACT

The focus of this study was to explore the suitability of trait mining using Focused Identification of Germplasm Strategy (FIGS) for the identification of resistance to Ug99. Ug99 is a new strain of the stem rust (*Puccinia graminis* Pers) fungal pathogen on wheat. The experiment was conducted as a "blind" study where the modeler calculating the trait mining model and predicting the scores for resistance to Ug99 did not know the actual trait scores. The predictions were validated against a dataset with the screening of wheat accessions against Ug99 conducted in Yemen during 2007. Only a small set with 20% of the Ug99 screening results were disclosed to the modeler for the training of the prediction model. The procedure followed was the same as a recent trait mining study on stem rust and designed to validate this approach in a "blind" study. The results from this study suggest that FIGS is well suited for the efficient sampling of genebank accessions for a target trait and useful for the identification of samples with a higher likelihood to hold resistance to Ug99.

Abbreviations: FIGS, focused identification of germplasm strategy; kNN, k-nearest neighbor; LR+, positive diagnostic likelihood ratio; PCA, principal component analysis; PPV, positive predictive value; PRESS, predicted residual sum of squares; SIMCA, soft independent method of class analogy; Ug99, stem rust race discovered in Uganda 1999.

INTRODUCTION

Focused Identification of Germplasm Strategy (FIGS) provides a sampling strategy to identify accessions from genebank collections for a target trait property (Mackay, 1986, 1990, 1995; Mackay and Street, 2004). The FIGS approach assumes a predictive link between the ecogeographic descriptions of the original collecting site for genebank accessions (with focus on landraces and crop wild relatives) and a target trait property such as agronomical factors or disease resistances. Recent studies have proposed the algorithms and methods for implementation of trait mining using FIGS (El Bouhssini et al., 2009, 2011; Endresen, 2010, Endresen et al., submitted [PAPER III]).

Stem rust is caused by the fungus *Puccinia graminis* (Pers) and has a long history as one of the most severe diseases on wheat (McIntosh, 1995). One of the most recent races of stem rust is called Ug99 and is exceptionally virulent. A report from CIMMYT (2005:9) reports yield loss up to 71% on experimental fields. Yield losses experienced in Kenya were reported at levels reaching 80% (KARI, 2005 *cf*. CIMMYT, 2005:10). Ug99 was discovered in Kalengyere Uganda in February 1999 (Pretorius, 2000) and immediately raised concerns because of the virulence to wheat plants carrying the stem rust resistance gene *Sr31*. This gene is one of the genes most widely used as protection against stem rust in modern cultivars (Wanyera et al., 2006). During the last decade Ug99 has caused an epidemic on wheat spreading through Eastern Africa turning north to enter Yemen in 2007 and most recently into Iran. Ug99 is likely to continue to spread and the identification of new sources of resistance effective to Ug99 are important to maintain the rational use of wheat in food production (Njau, 2010). A global collaborative initiative to fight this new epidemic of stem rust called the Borlaug Global Rust Initiative (www.globalrust.org, verified 30 Jan 2011) was established in 2008 (replacing the Global Rust Initiative from 2005).

MATERIALS AND METHODS

A set with 4563 genebank samples of bread wheat (*Triticum aestivum* L.) and durum wheat (*Triticum turgidum* L.) landraces were screened for resistance to race Ug99 of stem rust (*Puccinia graminis* Pers). The field trials were made in Yemen during the 2007 season. These screening results are not yet published in the scientific press. A recent study conducted by Endresen et al. (submitted [PAPER III]) explored prediction of stem rust resistance for wheat landraces using trait mining with FIGS. Stem rust data from the USDA GRIN system was split in two parts. The first part (training set) used to calibrate and tune a classification model; and the second part (test set) used to evaluate the predictive performance of the model in samples not yet exposed to the model. However the samples in the test set was known to the modeler and could unintentionally have influenced choices made during the preparation of the trait mining model.

To ensure the full absence of knowledge from the test set to influence the trait-mining model used for prediction of the trait scores, a new experiment for "blind" predictions were designed. Dr Kenneth Street representing ICARDA coordinated a follow-up experiment using the dataset with measurement of Ug99 resistance from Yemen 2007. We wanted to explore the performance of the trait mining models in a simulation of a real-life scenario where the trait scores predicted by the model was not known to neither the model nor to the modeler. A dataset including the accessions from the Ug99 trait dataset was prepared. This Ug99 set was split in 18% samples for a training set with trait scores included; and the remaining 82% as the test set without any of the trait scores included. The next task was to predict the trait scores for the accessions in the test set and report the predictions back to the project coordinator. For this experiment only the project coordinator knew the actual trait scores measured.

When genetic diversity from genebank collections are included in screening experiments to search for useful properties such as the resistance to pathogens, the final trait scores are obviously not known when selecting samples to include in the experiment. However the passport data including the geographic coordinates where the original material was collected are available. This study is thus a realistic simulation of the information available for the planning of germplasm evaluation experiments. In a real trait evaluation project

the accessions to be included would be selected more carefully than the random test set we used here. In a real life FIGS study the expert knowledge of the evaluator would likewise be utilized to complement the sampling by the computer model.

The Yemen dataset (stem rust, Ug99, 825 samples)

The Yemen dataset with accessions screened for Ug99 (2007) was matched with the accessions from the online SINGER (and USDA GRIN) databases to find more complete germplasm passport data. The dataset included a total of 4 563 accessions from a total of 1928 original collecting sites. Most of the samples were bread wheat (*Triticum aestivum* L.), but the dataset also included 114 accessions of durum wheat (*Triticum turgidum* L.) and 11 wheat samples of unidentified subspecies (genus *Triticum*). The trait observations for the training set with disclosed trait scores and 825 samples were reclassified to include three measurement levels. Trait scores reported as resistant "R" and medium resistant "MR" was classified as resistant samples (class 1). Trait scores reported as medium susceptible "MS" were assigned to class 2, and those reported as susceptible "S" as class 3. The location of the original collecting sites for the landraces in this set is illustrated by figure 1 and figure 2. The initial explorative principal component analysis (PCA) indicated that factorial analysis with decomposition of the climate data to principal components was a suitable approach for this dataset; the accessions were well separated in the score plots from the PCA analysis. This initial analysis did not indicate any outliers.

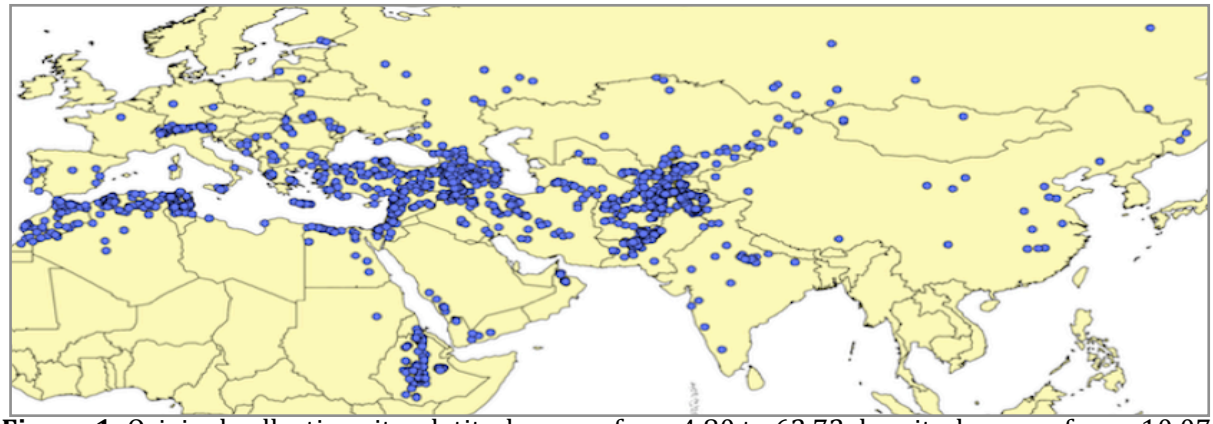


Figure 1: Original collecting sites, latitudes span from 4.80 to 62.72; longitudes span from -10.07 to 134.07. Complete set with a total of 4 563 accessions.

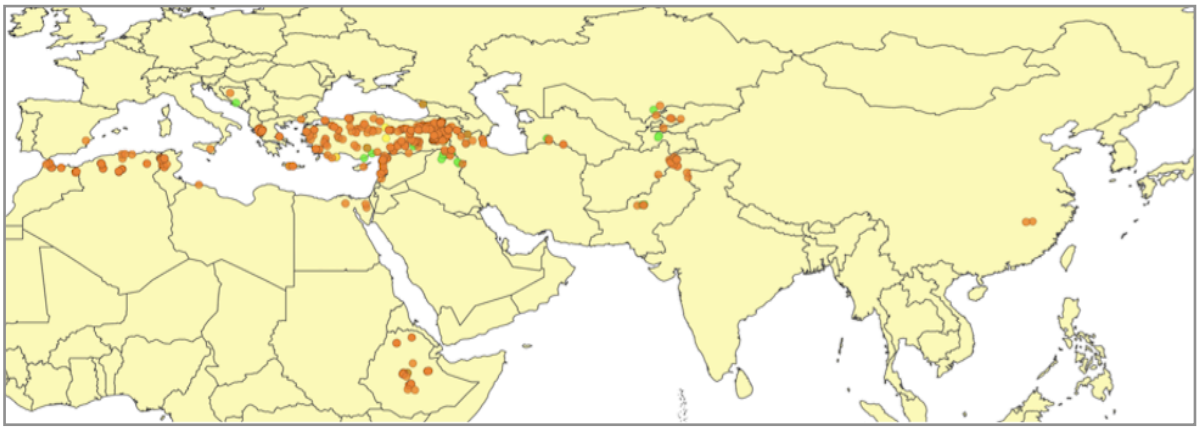


Figure 2: Training set from the Yemen Ug99 set (825 samples). With the disclosed actual observed trait scores from the Yemen Ug99 field trial, green for resistant samples, yellow for medium and red for susceptible samples.

ICARDA ecoclimatic database
Dr Eddy De Pauw (ICARDA GIS section) extracted the climate data for this study from the ecoclimatic information system maintained at ICARDA (De Pauw, 2003). The climate data was extracted using the longitude and latitude coordinates for the original collecting site for these landraces.

Prediction based on a model calibrated from USDA stem rust data
An initial prediction experiment was conducted using the same classification models as developed for a previous study (PAPER III). These models were calibrated with stem rust scores from field trials in Minnesota, USA during 1988 to 1994 (Bonman et al., 2007). However, the prediction of the Ug99 scores recorded in Yemen using this model performed surprisingly bad. The hit rate for resistant samples were only 9.55% while the ratio of resistant samples in the Ug99 set was reported by the project coordinator to be 10.2%. A second prediction with samples from the same collecting sites grouped together using the mean trait score for samples from the same site, resulted in an even lower hit rate.

Pre- study: Ug99 set (275+550 samples)
The next experiment was conducted including only the samples from the Ug99 set, and only the 825 samples with the trait scores reported. These samples were split in a training set with 550 samples (67%) and a test set with 275 samples (33%). The desktop study was in other aspects conducted as described by PAPER III, including the comparison of confidence intervals for the performance indicators calculated respectively for the accessions sampled by SIMCA and kNN, and for accessions randomly sampled.

Blind prediction of resistance to Ug99 (825+3738 samples)
The final experiment was the sampling of resistant samples from the 3738 samples from the Ug99 set with no trait scores disclosed to the modeler. The classification model was here calibrated using all of the 825 samples from the Ug99 set with trait scores reported. This trait-mining experiment was designed to simulate the sampling of accessions for a germplasm evaluation project with the (imaginary) capacity to screen 500 accessions. The task was thus to select 500 accessions from the Ug99 test set predicted to have a higher likelihood to be resistant to the Ug99 stem rust pathogen.

7 principal components were chosen as the SIMCA model complexity (figure 3). The predictions from the SIMCA model and the kNN model were thus combined with equal weight to form a so-called *"classifier ensemble"* (Kuncheva, 2004; Rokach, 2010). The top 500 accessions ranked by the predicted resistance to Ug99 were selected as the sampled subset with resistant samples (figure 7 and figure 8). These 500 accessions corresponds to 13.4% of the samples in the test set with 3738 samples, and thus slightly higher than the ratio resistant samples in the Ug99 set.

Classification methods (kNN and SIMCA)
The previous study by Endresen et al. (PAPER III) indicated that the SIMCA (soft independent modeling of class analogies; Wold, 1976; Wold and Sjostrom, 1977) and the kNN (k-nearest neighbor; Cover and Hart, 1967) were suited for the classification of stem rust trait scores. The results reported here was thus calculated using these two classification methods.

Evaluation of predictive performance

The target of this study was the identification of landraces predicted to be resistant to Ug99. Endresen et al. (PAPER III) reported that the positive predictive value (PPV) and the positive diagnostic likelihood ratio (LR+) were suitable performance indicators for trait mining predictions for stem rust, and thus selected for evaluation of the predictive classification results in this study.

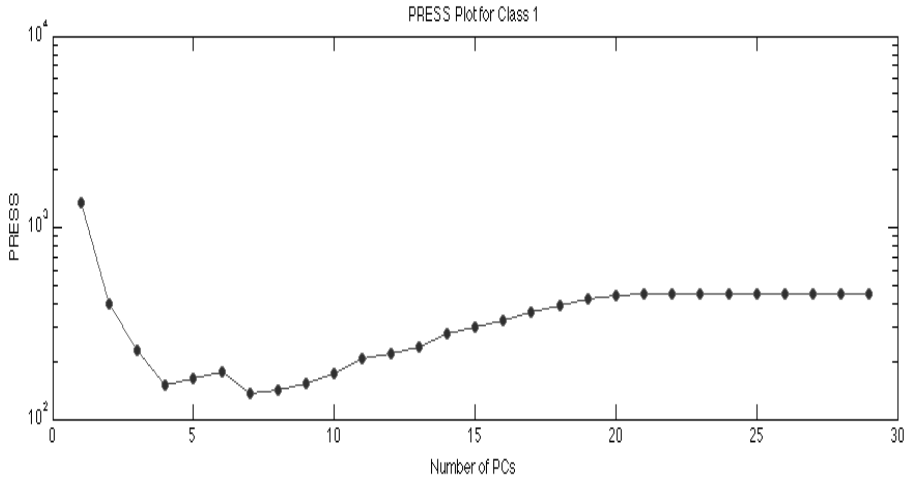
Software

The classification models were calculated using the PLS Toolbox (version 5.8; http://software.eigenvector.com/toolbox, verified 29 Jan 2011) for MATLAB (http://mathworks.com, verified 29 Jan 2011). The confidence interval for the performance indicators was calculated using the online Statistics Calculator from the Centre for Evidence Based Medicine (http://ktclearinghouse.ca/cebm/practise/ca/calculators/statscalc, verified 29 Jan 2011).

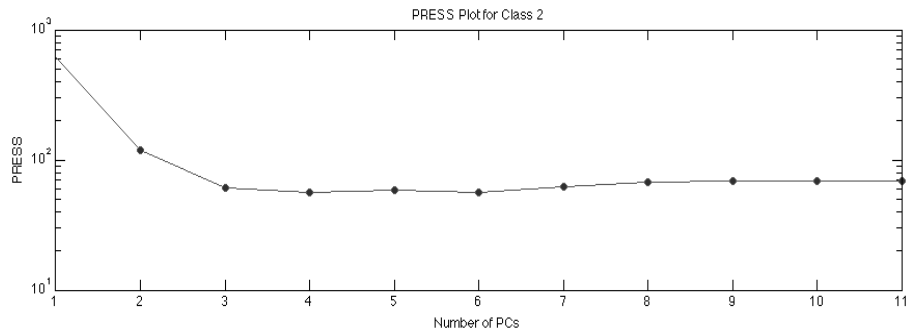
RESULTS

Pre- study: Ug99 set (275+555 samples)

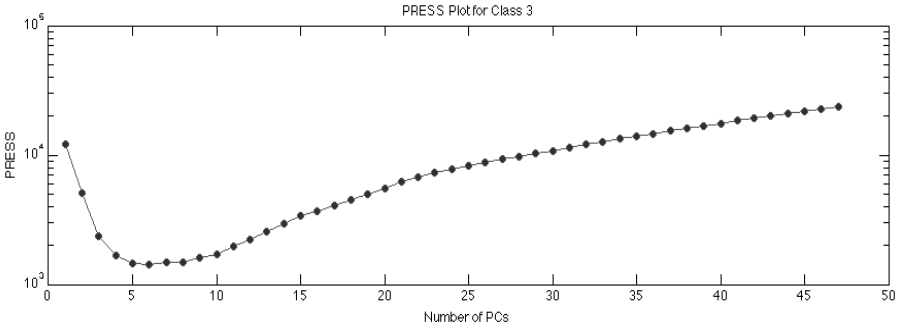
The plots in Figure provide the cross-validation results for the SIMCA model. The predicted residual sum of squares (PRESS) decreases towards a minimum around 7 principal components (PCs). The increase of PRESS after this level indicates that the SIMCA classifier starts to overfit the model to the data when more than 7 PCs are included in the model. Cross-validation when using all of the 825 samples with disclosed trait scores resulted in similar plots with the minimum PRESS around 7 PCs. The optimal model complexity of 7 principal components found here was thus selected as the preferred model complexity for the final SIMCA model for prediction of the 3738 "blind" samples.



(a)



(b)



(c)

Figure 3: Cross-validation results for the SIMCA model of the Yemen Ug99 set including 550 training set samples and 275 test samples. PRESS (prediction root mean square error) is plotted against the number of principal components for each of the PCA class models of the SIMCA model. (a): PCA model for class 1 with the resistant samples; (b) class 2 with intermediate resistance; and (c) class 3 with the susceptible samples.

The score plots (figure 4 and figure 5) indicate that the first principal component (figure 4) focus on the samples from Ethiopia. The second principal component has more equal focus on all samples. The third principal component (figure 5) focuses on the samples from Greece. Ethiopia is the country represented by the most samples (1260 samples), while Greece is the seventh country using this ranking approach (133 samples).

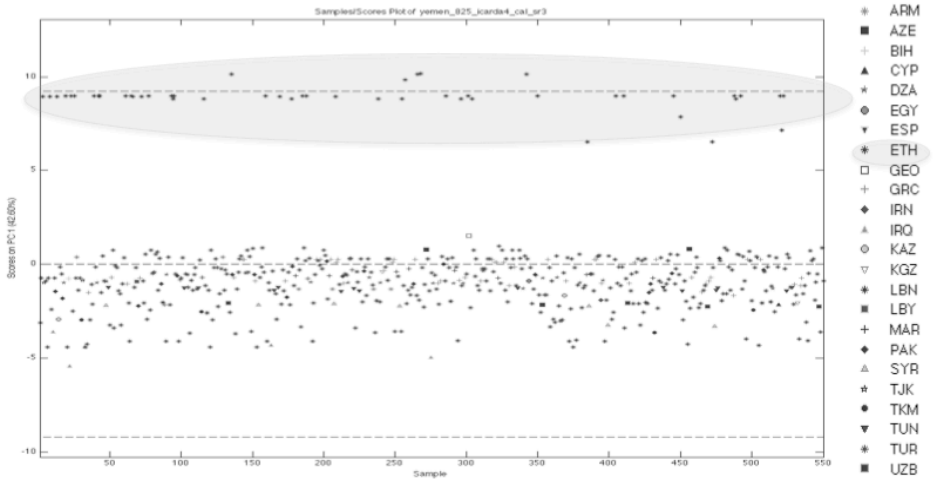


Figure 4: Score plot for the SIMCA model for class 1, first principal component. The samples from Ethiopia are highlighted.

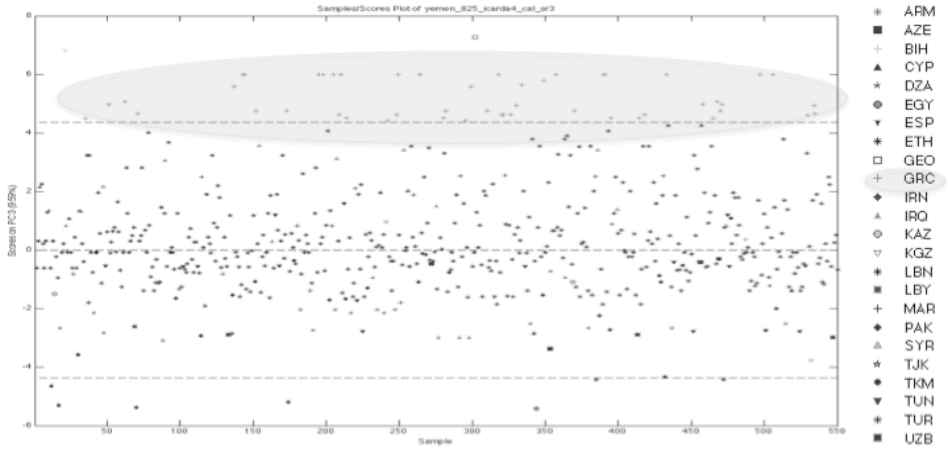


Figure 5: Score plot for the SIMCA model for class 1, principal component 3. The samples from Greece are highlighted.

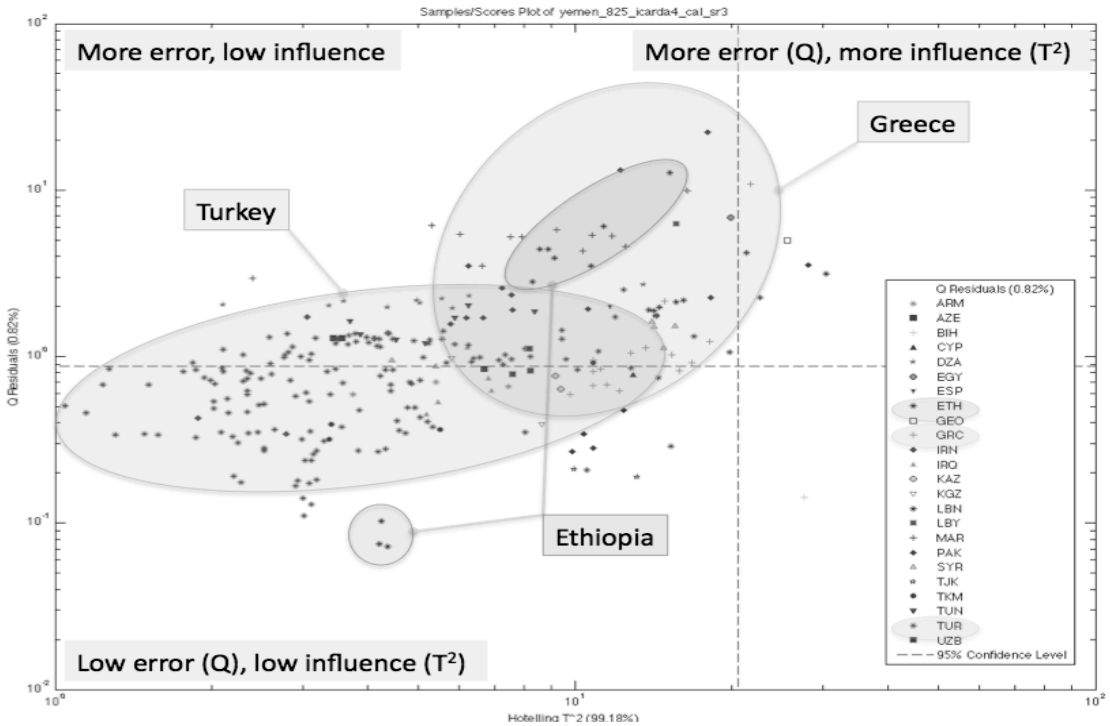


Figure 6: The influence plot (cross-validation) for the SIMCA model for class 1 (Yemen Ug99 set of 825 samples, random training set with 550 samples). The accessions in the lower left corner are well described by the model - the accessions in the upper right corner are less well described. Overall the model thus describes the dataset very well with none of the accessions separated from the other accessions.

The influence plot for the SIMCA model of class 1 (figure 6) indicates that the samples from Ethiopia are split in two groups with high and respectively low error (Q) and leverage (T^2). Most of the samples from Greece have high values for error and leverage. The samples originating from Turkey have generally low influence and error.

The performance indicators for these classification models are reported with Table 1. The hit rate for this trait mining experiment is higher than the hit rate reported by Endresen et al. (PAPER III) for a similar study using a stem rust dataset from USDA. Notice that the LR+ indicator should be used when comparing these results with the USDA stem rust results due to the different ratio resistant samples (prevalence) between these two datasets. This outcome was expected because the trait scores from the Ug99 set are recorded during one single trial season and at only one site, while the stem rust scores in the USDA set was a compilation of trait scores from 6 different seasons and two different experiment stations. The different screening conditions could thus introduce variation (genotype by environment interaction) in the trait scores that would be experienced as noise by the classification algorithm.

Notice that the 95% confidence interval for the SIMCA and kNN performance indicators overlap with the confidence interval for the random sampled accessions. There is thus no statistical significant support for a claim that these classification models perform better than the random sampling method. It is possible that the relatively broad confidence intervals are a result of too few samples in the test set. The correctly predicted resistant samples (true positives) were ranging from one to ten for the prediction models, and from zero to three for the random sampling method.

Table 1: Results from the evaluation of predictive performance in the 825 samples with trait score included (training set 550 samples and test set 275 samples)

Model	PPV	LR+
kNN	0.29 (0.13-0.53)	5.61 (2.21-14.28)
SIMCA (PC=11)	0.28 (0.14-0.48)	5.26 (2.51-11.01)
RANDOM	*0.06 (0.01-0.27)*	*0.95 (0.13-6.73)*

The 95% confidence interval for each of the performance indicators is included inside the parentheses.

Blind predictions for the Yemen set

The classification model for prediction of the 3728 unknown samples from the Ug99 set was calibrated using all of the 825 samples with revealed trait scores. The complete Ug99 set (4563 samples) was reported to have 10.2% resistant samples. The 500 samples selected by the kNN + SIMCA classifier ensemble was reported to the project coordinator who found that 129 samples (25.8%) were correctly predicted as resistant to Ug99. The proportion resistant samples in the set selected by the trait mining models were thus 2.5 times higher than the ratio resistant samples in the complete Ug99 set. The predicted Ug99 resistances for the 500 selected samples are included on a map with Figure 7, and for the entire Ug99 dataset with Figure 8.

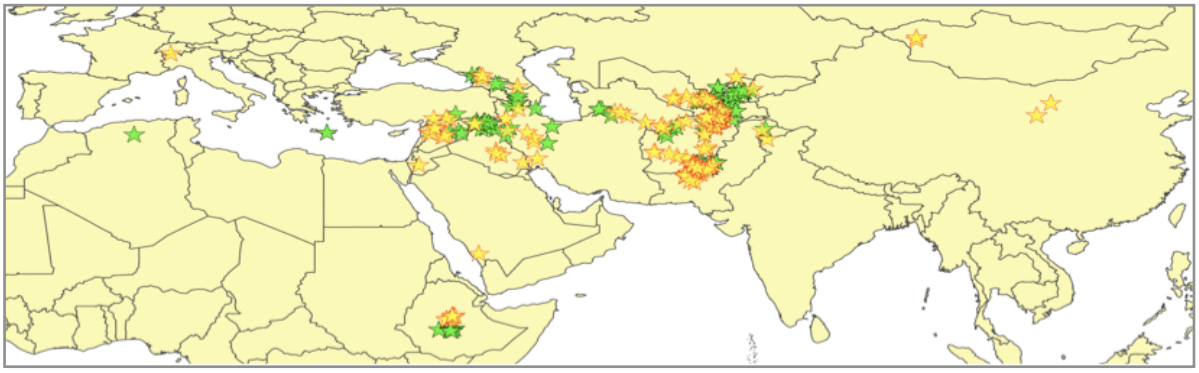


Figure 7: Predicted subset of 500 accessions. Green stars indicate accessions predicted as resistant by all of the trait mining models (SIMCA and kNN). Yellow stars indicate accessions predicted to be resistant only by some of the trait mining models. Number of accessions per country: PAK (149), AFG (109), ETH (89), UZB (31), TJK (28), IRQ (23), IRN (14), TKM (10), SYR (10), GEO (9), AZE (8), TUR (4), KAZ (3), KGZ (2), CHN (2), MNG (2), GRC (2), SAU (1), IND (1), CHE (1), DZA (1), JOR (1).

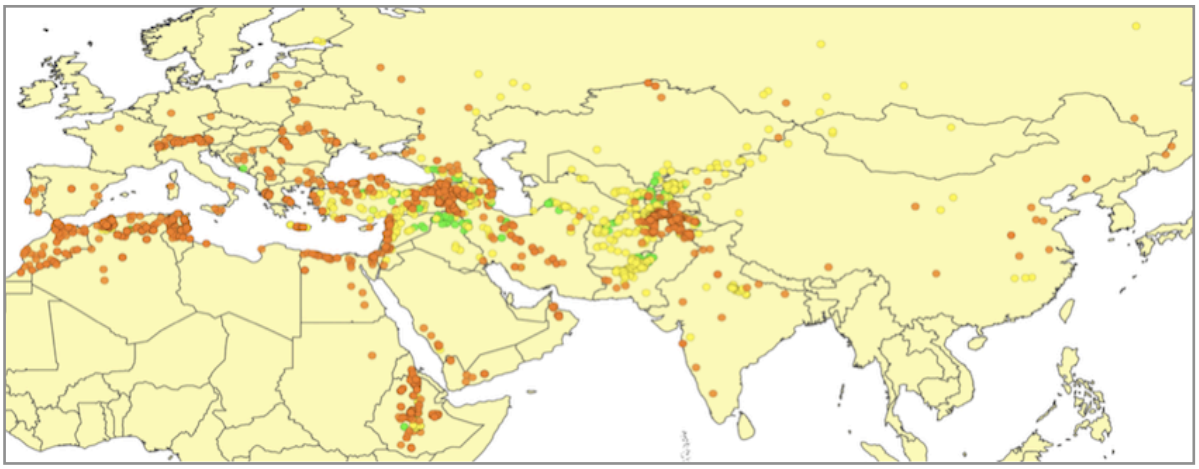


Figure 8: Samples from the Ug99 set with the predicted Ug99 resistance for the test set (3 738 samples). Green for predicted resistant samples, yellow for medium and red for susceptible samples. This is the ensemble prediction from kNN + SIMCA combined (mean).

DISCUSSION

From Minnesota to Yemen

The first predictions using the USDA stem rust set for model calibration, and the Ug99 set tested in Yemen, as the test set, did no better than chance. The Ug99 set included data for the screening of the Ug99 tribe of stem rust. This is another (and much more virulent) race than the type of stem rust screened for the USDA set. The USDA set was screened in Minnesota (Rosemount and St Paul), while the Ug99 set was screened in Yemen at different latitude and even more different longitude. The environmental conditions are very different in Yemen (dry and warm) compared to Minnesota (wet and colder). These differences in trial site could perhaps cause a difference in the expression of the stem rust resistance for these landrace genotypes.

Model structure

It is possible that there exists a predictive signal between the USDA stem rust set and the Ug99 set that these classifications models was unable to find. The classification model could perhaps be 'overfitted' in respect to the training data (USDA stem rust set). The classifier could also be too simple or have an inappropriate structure. Fuzzy samples are known to disturb the central covariance matrix of discriminant analysis models (Fielding, 2007). It is possible that a similar effect disturbed the SIMCA and kNN models in this study. Other methods such as Artificial Neural Networks (ANN; Bishop, 1996), Random Forest (Breiman, 2001) and other decision trees are less sensitive to such problems and might give a stronger predictive signal were other methods fail.

Unbalanced class

These stem rust datasets hold unequal proportions resistant samples compared to the much larger number of susceptible samples. The stem rust datasets have an asymmetric internal class data structure also called "unbalanced class". Different classifiers will be affected differently by the challenge of modeling very different class sizes. ANN has been reported to have fewer problems with asymmetrical internal class structures (Davies and Silverstein, 1995).

CONCLUSION

The results from this study suggest that trait mining with FIGS is a suitable approach to select samples for a specific trait such as the resistance to Ug99.

LITERATURE

Altman, D.G., and J.M. Bland. 1994b. Statistical notes: Diagnostic tests 2: predictive values. BMJ 309(6947): 102. Available at http://www.bmj.com/content/309/6947/102.1.full, verified 30 Jan 2011.

Agresti, A. (2002). Categorical data analysis. Second edition. John Wiley and Sons, Hoboken, New Jersey, USA. ISBN: 9780471360933.

Barker, M., and W. Rayens. (2003). Partial least squares for classification. Journal of Chemometrics 17(3): 166-173. DOI: 10.1002/cem.785.

Bishop, C. (1996). Neural networks for pattern recognition. Oxford University Press, UK. ISBN: 978-0198538646.

Breiman, L. (2001). Random forests. Machine Learning 45(1): 5–32. DOI: 10.1023/A:1010933404324.

Bonman, J.M., H.E. Bockelman, Y. Jin, R.J. Hijmans, and A.I.N. Gironella (2007). Geographic distribution of Stem rust resistance in wheat landraces. Crop Science 47: 1955-1963. DOI: 10.2135/cropsci2007.01.0028.

Borlaug Global Rust Initiative. Available at http://www.globalrust.org, verified 30 Jan 2011.

CIMMYT (2005). Sounding the alarm on global Stem rust. An assessment of race Ug99 in Kenya and Ethiopia and the neighboring regions and beyond, by the expert panel on the Stem rust outbreak in Eastern Africa, 29 May 2005. CIMMYT, Mexico. Available at http://www.globalrust.org/db/attachments/about/2/1/Sounding the Alarm on Global Stem Rust.pdf, verified 30 Jan 2011.

Cover T.M., P.E. Hart (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1): 21-27.

Davies, P., and B.R. Silverstein (1995). A comparison of neural nets to statistical stubborn classification problems. *In*: International Conference on Acoustics, Speech, and Signal Processing. ICASSP-95. Detroit, MI, USA. DOI: 10.1109/ICASSP.1995.479732.

De Pauw, E. (2008). Climatic and soil datasets for the ICARDA wheat genetic resource collections of the Eurasia region. Explanatory Notes. ICARDA GIS Unit, Aleppo, Syria. 68 p. Available online at http://geonet.icarda.cgiar.org/geonetwork/data/regional/GRU_NetBlotch/Doc/Report_NetBlotch.pdf, 6.6 MB, verified 30 Jan 2011.

El 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.

El Bouhssini, M., K. Street, A. Amri, M. Mackay, F.C. Ogbonnaya, A. Omran, O. Abdalla, M. Baum, A. Dabbous, and F. Rihawi (2011). Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS). Plant Breeding 130(1): 96-97. DOI: 10.1111/j.1439-0523.2010.01814.x

Endresen, D.T.F. (2010). Predictive association between trait data and ecogeographic data for Nordic barley landraces. Crop Science 50(6): 2418-2430. DOI: 10.2135/cropsci2010.03.0174.

Fielding, A.H. (2007). Cluster and classification techniques for the biosciences. Cambridge University Press, Cambridge, UK. ISBN-13 978-0-521-85281-4.

Fisher, R.A. (1936). The use of multiple measurements in taxonomic problems. Ann. Eugen. 7: 179-188.

Gwet, K.L. (2010). Handbook of inter-rater reliability (Second edition), The definitive guide to measuring the extent of agreement among multiple raters. Advanced Analytics, LLC, Gaithersburg, MD, USA. 208 p. ISBN: 9780970806222.

Hawkins, D.M. (2004). The problem of overfitting. Journal of Chemical Information and Computer Sciences 44: 1-12. DOI**:** 10.1021/ci0342472.

Hawkins, D.M., S.C. Basak, and D. Mills (2003). Assessing model fit by cross-validation. Journal of Chemical Information and Computer Sciences 43(2): 579–586. DOI: 10.1021/ci025626i.

KARI (2005). Effect of a new race on wheat production/use of fungicides and its cost in large vs small scale farmers, situation of current cultivars. Kenya Agricultural Research Institute (KARI), Njoro, Kenya. (*cf*. CIMMYT, 2005:10)

Kohavi, R., and F. Provost (1998). Glossary of terms. Machine Learning 30(2-3): 271-274. DOI: 10.1023/A:1017181826899.

Krzanowski, W. J. (1988). Principles of multivariate analysis: A user's perspective. Oxford University Press, Oxfordshire, UK. 563 p. ISBN: 0198522118.

Kuncheva, L.I. (2004). Combining pattern classifiers. Methods and algorithms. John Wiley & Sons, Hoboken, New Jersey, USA. ISBN: 0-471-21078-1.

Mackay, M.C. (1986). Utilizing wheat genetic resources in Australia. p. 56-61. *In*: McLean, R. (ed). Proceedings of the 5th Assembly Wheat Breed Wheat Society in Australia, Merredin 18-22 Aug 1986. Western Australian Department of Agriculture, Perth, Australia. 580 p. ISBN: 9780730913269.

Mackay, M.C. (1990). Strategic planning for effective evaluation of plant germplasm. p. 21–25. *In*: Srivastava J.P., and A.B. Damania (eds). Wheat genetic resources: Meeting diverse needs. John Wiley & Sons, Chichester, UK. ISBN 0-471-92880-1.

Mackay, M.C. (1995). One core collection or many? p. 199-210. *In*: Hodgkin T., A.H.D. Brown, Th.J.L. van Hintum, and A.A.V. Morales (eds). Core collections of plant genetic resources. John Wiley & Sons, Chichester, UK. ISBN: 471-95545-0.

Mackay M.C. and K. Street (2004). Focused identification of germplasm strategy – FIGS. p. 138-141. *In*: Black, C.K., J.F. Panozzo, and G.J. Rebetzke (eds). Proceedings of the 54th Australian Cereal Chemistry Conference and the 11th Wheat Breeders' Assembly. Royal Australian Chemical Institute, Melbourne, Australia.

McIntosh, R.A., C.R. Wellings, and R.F. Park (1995). Wheat rusts: An atlas of resistance genes. CSIRO, Melbourne, Victoria, Australia. ISBN: 0-643-05428-6.

Pretorius, Z.A., R.P. Singh, W.W. Wagoire, and T.S. Payne (2000). Detection of virulence to wheat stem rust resistance gene *Sr31* in *Puccinia graminis* f. sp. *tritici* in Uganda. Plant Disease 84(2): 203. DOI: 10.1094/PDIS.2000.84.2.203B.

Rokach, L. (2010). Pattern classification using ensemble methods. Series in machine perception and artificial intelligence - Vol. 75. World Scientific Publishing Co. Pte. Ltd., Singapore. ISBN: 9814271063.

Stevens, S.S. (1946). On the theory of scales of measurement. Science 103(2684): 677-680. DOI: 10.1126/science.103.2684.677.

Wanyera, R., M.G. Kinyua, Y. Jin, and R.P. Singh (2006). The spread of Stem rust caused by *Puccinia graminis* f. sp. *tritici* virulence on *Sr31* in wheat in Eastern Africa. Plant disease 90(1): 113. DOI: 10.1094/PD-90-0113A.

Wold, H. (1966). Estimation of principal components and related models by iterative least squares. p. 391-420. *In*: Krishnaiaah P.R. (ed). Multivariate Analysis. Academic Press, New York, USA.

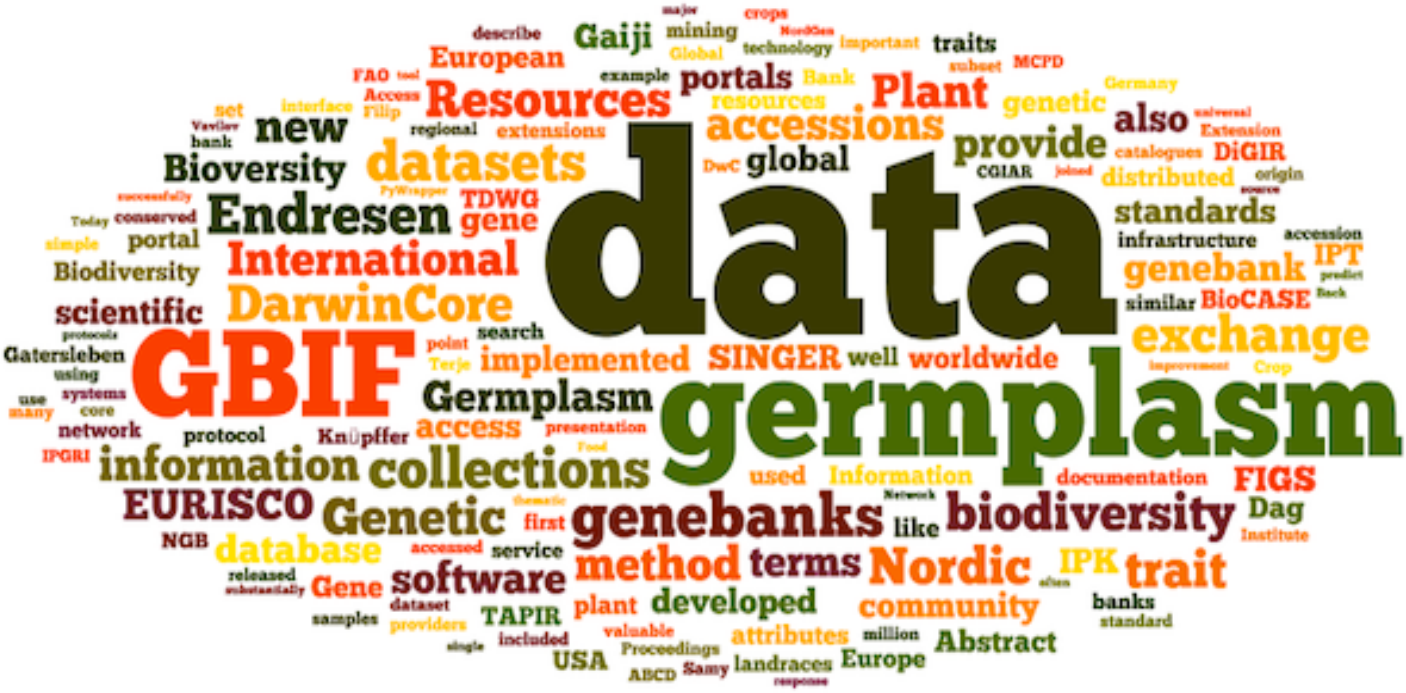
Wold, S. (1976). Pattern recognition by means of disjoint principal component models. Pattern Recognition 8: 127-139.

Wold, S., and M. Sjostrom (1977). SIMCA: A method for analyzing chemical data in terms of similarity and analogy. p. 243-282. *In*: Kowalski, B.R. (ed). Chemometrics Theory and Application, American Chemical Society Symposium Series 52. American Chemical Society Washington D.C., USA.

Wold, S., A. Ruhe, H. Wold, and W.J. Dunn (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM Journal of Scientific and Statistical Computing 5: 735-743.

ABSTRACTS

- **ABSTRACT 1**: Endresen, D.T.F. and B. Skovmand (2006). Trait Mining in Gene Banks.
- **ABSTRACT 2**: Endresen, D.T.F., J. Bäckman, H. Knüpffer, and S. Gaiji (2006). Exchange of Germplasm Dataset with PyWrapper/BioCASE. *Peer review*
- **ABSTRACT 3**: Knüpffer, H., D.T.F. Endresen, and S. Gaiji (2007). Integrating Genebanks Into Biodiversity Information Networks. *Peer review*
- **ABSTRACT 4**: Endresen, D.T. (2008). Biodiversity data exchange software, hands-on exercises with TAPIR software.
- **ABSTRACT 5**: Endresen, D., S. Gaiji, and T. Robertson (2009). Darwin Core Germplasm Extension and deployment in the GBIF infrastructure. *Peer review*
- **ABSTRACT 6**: Endresen, D. (2010). A Lifeboat to the Gene Pool. Predictive association between trait data and eco-geographic data for identification of trait properties useful for improvement of food crops.
- **ABSTRACT 7**: Gaiji, S., D. Endresen, J. Nordling, S. Dias, E. Arnaud (2010). Beyond Darwin Core: Challenges in mobilizing richer content. *Peer review*



Trait Mining in Gene Banks.

Dag Terje Filip Endresen and Bent Skovmand. Nordic Gene Bank, P.O. Box 41, Alnarp, SE-230 53, SWEDEN

Gene banks have search catalogues where potential users can search for interesting germplasm. These germplasm collections have most of their attributes in common, but the terminology used to describe them may differ substantially. Standards for data exchange and data integration have been developed and partly implemented by gene banks. Today no single point of access to search germplasm worldwide across all gene bank collections exists. The task of combining germplasm documentation datasets from different sources is challenging and methods to improve data interoperability are needed. The NGB (Nordic region), EURISCO (Europe), SINGER (CGIAR) and GRIN (USA) catalogues provide access to the passport data of major collections. But these catalogues only integrate a small set of common descriptors describing the accession IDs and origin of the accessions. Descriptive traits are often excluded and the individual online gene bank catalogues needs to be visited to find this data type. A lack of a user-friendly access to the relevant documentation on the conserved accessions is a major constraint for a wider use of the germplasm holdings in gene banks. The Nordic Gene Bank has started a project on the development of methodologies and tools for trait and gene mining in the existing germplasm data on gene bank accessions. Longitude and latitude of the germplasm site origin together with measured traits and associated environmental attributes on climate, abiotic and biotic stresses will be used to build a habitat signature or digital pattern. This pattern can then be applied to germplasm passport data from other gene banks or collections of genetic resources to predict unknown trait characters. A similar approach is today in use to predict species distribution, but these methods have not yet been applied to trait mining for passport data on gene bank collections of germplasm.

Wednesday, November 15, 2006 273-10
Conservation and Evaluation of Crop Germplasm
C08 Plant Genetic Resources
The ASA-CSSA-SSSA International Annual Meetings (November 12-16, 2006)

Endresen, D.T.F. and B. Skovmand (2006). Trait Mining in Gene Banks. American Society of Agriculture, ASA-CSSA- SSSA, Indianapolis, USA, 12-16 November 2006. Available at http://a-c-s.confex.com/crops/2006am/techprogram/P26713.HTM

Proceedings of TDWG, 2006

Exchange of Germplasm Datasets with PyWrapper/BioCASE

Dag T. F. Endresen, Johan Bäckman, Helmut Knüpffer, Samy Gaiji

There are more than six million ex situ germplasm accessions of agricultural and horticultural crops conserved worldwide by genebanks (seed banks), according to the FAO. These germplasm collections share most of their attributes, but database systems and data models implemented may differ substantially. The International Plant Genetic Resources Institute, IPGRI, has developed standards for data exchange and data integration, which are implemented by many genebanks. Germplasm collections also share many attributes with other biodiversity collections, such as natural history museums, botanical gardens or herbaria. Today there is no single point of access allowing discovering germplasm samples across all genebank collections worldwide. Germplasm data portals like EURISCO (European genebanks), SINGER (CGIAR genebanks), USDA-GRIN (USA) and NGB (Northern Europe) successfully demonstrate that distributed data on germplasm accessions (genebank seed samples) can be mapped to common standards and thus accessed from global and regional data portals. These regional portals have so far been implemented as classical data warehouses. The attributes of the source datasets have been transformed to the agreed data exchange standard and included in a central database or index.

GBIF supports data flow from simple web services implemented by DiGIR or BioCASE/PyWrapper data provider software installed locally at each data source node. Such wrappers can be implemented for different database systems and do not require modification of the local database structure. Any update of contents of the local database will immediately be visible for search portals. A number of genebanks have already joined GBIF as data providers. This process was initiated by IPK Gatersleben, Germany. The first genebank to provide its accession data records to GBIF was the Nordic Gene Bank (North Europe) in March 2004. IHAR (Poland) and IPK Gatersleben (Germany) followed soon after. Later also USDA-GRIN (USA, 2005) and WUR, CGN Wageningen (The Netherlands, 2006) became GBIF data providers. The CGIAR genebanks through SINGER, and EURISCO representing most European genebanks, have also joined GBIF (2006) and provide data records to the GBIF index. The GBIF data portal provides a new and valuable channel to promote the germplasm datasets. The adoption of the PyWrapper software has proven relatively simple, and a genebank providing data to GBIF will, with very small extra efforts, be able to provide the same dataset to EURISCO or SINGER, using the same data standards. Exchange of germplasm data with PyWrapper has successfully been tested with the ABCD, Darwin Core, GCP Passport, and the MCPD data standards. Work is in progress to further implement PyWrapper as the preferred data exchange tool for the genebanks providing data to EURISCO and SINGER. Development of the new TAPIR protocol will soon provide many important and promising improvements to the data harvesting and indexing routines of germplasm data portals. Improved data harvesting routines and a time-to-live attribute for datasets and individual records are also under development. The TDWG standards on GUIDs will also play an important role.

Some germplasm data portals:
EURISCO http://eurisco.ecpgr.org/ Europe
SINGER http://singer.grinfo.net/ CGIAR
USDA-GRIN http://www.ars-grin.gov/ USA
SESTO http://www.ngb.se/sesto/ Nordic Countries
FAO WIEWS http://apps3.fao.org/wiews/

Endresen, D.T.F., J. Bäckman, H. Knüpffer, and S. Gaiji (2006). Exchange of germplasm dataset with PyWrapper/BioCASE. p. 8. *In*: Belbin, L., A. Rissoné, and A. Weitzman (eds). Proceedings of TDWG 2006. 16 October 2006, Missouri Botanical Garden in St. Louis, Missouri, USA. ISBN: 1-930723-56-3. Available at http://www.tdwg.org/proceedings/article/view/64 [Peer review]

INTEGRATING GENEBANKS INTO BIODIVERSITY INFORMATION NETWORKS

Helmut Knüpffer[1], Dag Terje Filip Endresen[2] and Samy Gaiji[3]
[1]Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), D-06466 Gatersleben, Germany; [2]Nordic Gene Bank, POB 41, SE-230 53 Alnarp, Sweden; [3]Bioversity International, Via dei Tre Denari 472/a, I-00057 Maccarese, Rome, Italy; (email: knupffer@ipk-gatersleben.de)

As called by the International treaty on PGRFA (Article 17), existing information systems should collaborate to form a global information network. According to the Food and Agriculture Organization of the United Nations (FAO), there are more than six million *ex situ* germplasm accessions of agricultural and horticultural crops conserved by genebanks worldwide. Many genebanks have computerised their information, but the database software and data models implemented may differ substantially between genebanks. Under the coordination of Bioversity International (formerly IPGRI), standards for the exchange and integration of germplasm information were developed and adopted by many genebanks.

Today there is no single point of access to all genebank collections worldwide at the accession level, but germplasm data portals such as the EURISCO (European search portal), numerous Central Crop Databases (CCDBs), the Nordic Genebank (NGB, Northern Europe), and the CGIAR's System-wide Information Network for Genetic Resources (SINGER) among others, show that distributed data on genebank accessions can be accessed from global and regional as well as crop-specific data portals, implemented as classical data warehouses.

The Global Biodiversity Information Facility (GBIF) promotes the exchange of biodiversity related information using a new information technology called web services. Such technology deployed at the level of data providers offers the opportunity to tap remotely into the "living" database. Germplasm collections are very similar information-wise to other biodiversity collections, such as natural history museums, botanical gardens or herbaria. Initiated by Bioversity, GBIF data exchange technology was further developed to suit the needs of the PGR community.

From 2004 on, several genebanks became GBIF data providers, the first being NGB, IHAR (Poland) and IPK (Germany), followed by USDA- GRIN (USA) and CGN (Netherlands). Bioversity joined GBIF in 2006, bringing in SINGER and EURISCO. Thus, with two million accessions, one- third of the world's germplasm holdings are presently searchable via GBIF. The wide adoption of GBIF technology for PGR information exchange would facilitate an alliance of distributed germplasm information systems. Already, Bioversity International has undertaken a feasibility study for such a global system comprising more than 2.3 million accessions.

Knüpffer, H., D.T.F. Endresen, and S. Gaiji (2007). Integrating Genebanks Into Biodiversity Information Networks. Pages 34-35. *In*: Proceedings from the 18th EUCARPIA Genetic Resources Section Meeting. 23-26 May 2007, Piešťany, Slovak Republic. ISBN: 9788088872634. URL: http://vurveucarpia.kios.sk/abstracts/1/ [*Peer review*]

Biodiversity data exchange software, hands-on exercises with TAPIR software

Dag Terje Filip Endresen**,** Nordic Genetic Resources Center, and Bioversity International

A standardized data exchange protocol is a central part of implementing a biodiversity data network. A growing number of national and global projects link up with the same distributed datasets, and there is thus an increasing demand for generic and globally agreed data exchange protocols. Participation and the sharing of your institute datasets with global and national biodiversity projects is important for your public and scientific visibility, promoting the use of your data and ultimately for the continued funding of your institutional activities. The XML (Extensible Markup Language) is a major step forward to structure your data and to make them available in a universal format. Web service standards from W3C (World Wide Web Consortium) have come a long way to standardize the data transfer protocol. Implementation of XML and W3C standards guarantee that a recipient of your data can access and read your data. But it is no guarantee that the recipient will understand the data presented. Most primary biodiversity datasets have data units with a scientific name (nomenclature), a local id number and geospatial data for the gathering or observation site. Common concepts like these are added to a shared standard schema. Current concept schemas for primary biodiversity data include the Darwin Core (DwC) and the ABCD (Access to Biological Collections Data).

The DiGIR (2002) and BioCASE (2003) database wrappers provides you with a user- friendly software tool to map your database to (in particular) the Darwin Core and ABCD schema; and for setting up an online data provider service. The DiGIR and BioCASE software will wrap your datasets as a standardized bobble with a universal interface. There are simple rules to define how a service request looks like, and simple rules to define how the service response will look like. DiGIR and BioCASE have different request and response formats (interface protocols). In 2004 the TDWG (Biodiversity Information Standards) consortium initiated the work on TAPIR (TDWG Access Protocol for Information Retrieval) to unify and expand the features and interface protocol for DiGIR, BioCASE. PyWrapper3 (based on BioCASE, Python) was the first TAPIR software implementation and released in late 2006. TapirLink (based on DiGIR, PHP) was released shortly after in early 2007. There are also TAPIR implementations for the .NET framework and Java (under development). Independent on which of these software implementations you choose for your dataset, each service will have the same universal TAPIR protocol interface and can be accessed without the need to know which underlying software is used.

Relevant URLs
http://www.tdwg.org/activities/tapir/
http://www.pywrapper.org/
http://wiki.tdwg.org/twiki/bin/view/TAPIR/TapirLink

Endresen, D.T. (2008). Biodiversity data exchange software, hands-on exercises with TAPIR software. 3 Dec 2008, Stockholm Biodiversity Informatics Symposium 2008 (SBIS2008), Swedish Museum of Natural History, Stockholm, Sweden. Available at http://artedi.nrm.se/sbis2008/

Proceedings of TDWG, 2009

DarwinCore Germplasm Extension and deployment in the GBIF infrastructure

Dag Endresen, Samy Gaiji, and Tim Robertson

DarwinCore is designed around a set of general terms applicable for most biodiversity datasets. DarwinCore also implements a model of extensions to the core terms, designed to include terms of more specific utility in thematic domains. The DarwinCore Germplasm Extension has been developed to include the additional terms required to describe germplasm samples maintained by genebanks worldwide. The most widely used terms to describe germplasm samples are included in the Multi-Crop Passport Descriptors (MCPD) developed and published in December 2001 by Bioversity International (formerly IPGRI) and the Food and Agriculture Organization of the United Nations (FAO). In 2005 the MCPD terms were integrated to the ABCD standard (Access to Biological Collections Data). This work paved the way for the implementation of the BioCASE data publishing toolkit in the plant genetic resources community, and the improved sharing of germplasm datasets within the community as well as with the GBIF Network. The DarwinCore Germplasm Extension includes in a similar manner the missing terms from the MCPD standard. A few additional terms were included for description of the germplasm in relation to the new international Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) and other regulatory mechanisms. You will also find included the new terms for exchange of germplasm trait measurements developed in Europe for the ECPGR network (European Cooperative Programme for Plant Genetic Resources).

The GBIF Integrated Publishing Toolkit (IPT) was released in March 2009. In a similar manner as for the DarwinCore extensions, the IPT has the great advantage to be extended with more domain specific application schemas or extensions. When publishing a dataset with the IPT, the publisher can select terms from the general DarwinCore as well as from the extensions. The GBIF Harvesting and Indexing Toolkit (HIT) to be released in October 2009 will further make the indexing of distributed and heterogeneous datasets easier for Network managers. IPT also introduces the new DarwinCore archive. The DarwinCore archive will significantly speed up the indexing of datasets by a central portal like the GBIF portal or a thematic portal like for example the new global germplasm portal (Global-ALIS). Both the IPT and the HIT are able to synchronize with the GBIF Global Biodiversity Resources Discovery System (GBRDS). Thus datasets, data sharing protocols or extensions registered at the GBRDS can be easily discovered and accessed by a distributed thematic or regional biodiversity information network or for example by a more specific data analysis tool.

DarwinCore: http://rs.tdwg.org/dwc/terms/index.htm
Germplasm Extension: http://rs.nordgen.org/dwc/
GBIF IPT: http://code.google.com/p/gbif-providertoolkit/
GBIF HIT: http://code.google.com/p/gbif-indexingtoolkit/
GBIF Portal: http://data.gbif.org/ and http://www.gbif.org
Prototype Global ALIS: http://www.global-alis.org/

Endresen, D., S. Gaiji, and T. Robertson (2009). Darwin Core germplasm extension and deployment in the GBIF infrastructure. p. 78. *In*: Weitzman, A.L. (ed). Proceedings of TDWG 2009. 12 Nov 2009, Montpellier, France. Available at http://www.tdwg.org/proceedings/article/view/464 [Peer review]

Predictive association between trait data and eco-geographic data for identification of trait properties useful for improvement of food crops

Dag Terje Filip Endresen, Nordic Genetic Resources Center (NordGen)

Focused Identification of Germplasm (FIGS) is a new method to select plant genetic resources for the improvement of food crops. Traditional cultivars (landraces) and crop wild relatives (CWR) provide a valuable source for novel alleles in crop improvement programs, but conserved landraces and CWR often lack important documentation. Genebank collections worldwide provide access to plant genetic resources including online documentation. However incomplete documentation, and in particular the lack of relevant characterization and evaluation data (traits), often limit the efficient use of plant genetic resources. This presentation will demonstrate how trait mining with the new FIGS method can be used to predict missing trait information for barley landraces. The difference between the FIGS method and the core collection strategy is a good starting point to describe the new FIGS method. The core collection strategy aims at building a smaller subset from a larger set of germplasm accessions while keeping the subset as representative as possible for the complete set. The FIGS method will in a similar manner produce a smaller subset from a larger set of germplasm accessions, but the FIGS method will aim to target the smaller subset to a specific desired trait expression. The FIGS method assumes a predictive association between the trait and the eco-geographic attributes of the location of origin for the germplasm accession (landraces). The presentation will further give an introduction to the trait-mining computer modeling approach. For the first practical example, eco-geographic data from the location of origin for 14 Nordic landraces of barley (*Hordeum vulgare* L.) was successfully correlated to morphological traits using a modern multilinear data modeling method (N-PLS). The second example is a FIGS dataset on resistance to net blotch (*Pyrenophora teres* Drechs.) in barley. These two examples suggests that trait mining can efficiently be used as a targeted germplasm selection method and complement or replace the current core selection method in situations when the requirements for the trait mining method is fulfilled.

Vavilov Seminars: The lectures series was established as a platform for presentations in the field of research on cultivated plants, especially plant genetic resources, in 1991 together with *Vavilov Evenings*. Focus is the presentation of scientific topics in the plant genetic field.
Waterman Seminars: This seminar series is intended to provide a scientific stage for outstanding internal progress reports as well as presentations of national and international guests of the IPK on the field of Bioinformatics. Moreover it serves as a platform to join the Bioinformatics at the IPK as well as to help integrating it with all other departments.
URL: http://www.ipk-gatersleben.de/Internet/Veranstaltungen/KolloquienSeminare

Endresen, D. (2010). A Lifeboat to the Gene Pool. Predictive association between trait data and eco-geographic data for identification of trait properties useful for improvement of food crops. Vavilov and Waterman seminar. 12 May 2010, Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany. Available at http://bioinformatics.ipk-gatersleben.de/seminars/waterman/announcement_endresen.pdf

Proceedings of TDWG, 2010

5.3 Beyond DarwinCore: Challenges in mobilizing richer content

Samy Gaiji, Dag Endresen, Jonas Nordling, Sonia Dias, and Elizabeth Arnaud

One of the challenges of the scientific community is access to richer biodiversity content than the DarwinCore (DwC) concepts in order to span to phenotypic as well as genomics and ecosystems domains. The further engagement of the scientific community requires the biodiversity informatics community to provide the infrastructure response to such fundamental need. In 2009, the genebank community developed its first enriched extension to the new DwC version covering the description of phenotypic traits. In 2010, within the strategy of GBIF to expand its global infrastructure a feasibility study was initiated with NordGen and Bioversity International to assess the scalability of the GBIF infrastructure in meeting the needs of the European genebank community. Various installations of the GBIF Integrated Publishing Toolkit (IPT) were deployed within the European Plant Genetic Resources Catalogue (EURISCO) network of publishers using the DwC genebank extension.

Access to such richer biodiversity information through the European Plant Genetic Resources Catalogue (EURISCO) is critical to the scientific and policy-making users (e.g., identification of the most valuable germplasm with economically valuable traits). This presentation will highlight the lessons learnt of this feasibility study and provide recommendations on ways forward for the expansion of the GBIF infrastructure in support of scientific communities.

Gaiji, S., D. Endresen, J. Nordling, S. Dias, and E. Arnaud (2010). Beyond Darwin Core: Challenges in mobilizing richer content. p. 15-16. *In*: Weitzman, A.L. (ed). Proceedings of TDWG 2010. 28 Sep 2010, Woods Hole, Massachusetts, USA. Available at http://www.tdwg.org/fileadmin/2010conference/documents/Provisional_Proceedings_of_TDWG_2010.pdf [*Peer review*]

Appendix 1: Abbreviations

Abbreviations:

ABCD	Access to biological collections data (2005)
ABS	Access and benefit sharing [legislation under the CBD]
ANN	Artificial neural networks
API	Application programming interface
ASCII	American standard code for information interchange
BC	Before Christ, (AD, from Latin: *Anno Domini*)
BIOCLIM	Bioclimatic (ecological niche modeling approach)
BIRS	Biological information retrieval system (1983-1989)
BP	Before present (*'Before Physics'*, before 1950)
C&E	Characterization and evaluation data
CD-ROM	Compact disc, read-only medium
cf.	confer (from Latin: *conferre*)
COMECON	Council for mutual economic assistance (Russian: *Совет экономической взаимопомощи, Sovet ekonomicheskoy vsaymopomoshchi, СЭВ, SEV*) (1949-1991)
CWR	Crop wild relative
CWRIS	Crop wild relative information system (2005)
CWRML	Crop wild relative markup language (2005)
DNA	Deoxyribonucleic acid (double helix, 1953)
DOI	Digital object identifier
DwC	Darwin Core
DwC-germplasm	Darwin Core extension for genebanks
ECCDB	European central crop database
EURISCO	European catalogue of *ex situ* genebank collections (Ancient Greek: I find, I discover) (2003)
EXIR	Executive information retrieval (1975)
FIGS	Focused identification of germplasm strategy (2004) [1986]
GARP	Genetic algorithm for rule-set production (2003)
GATT	General agreement on tariffs and trade (1949-1993)
GCPCR	Generation challenge programme (GCP) Central registry
GeneSys	Genetic resources information system
GIS	Geographic information system
GLM	General linear model
GMO	Genetically modified organism
GPA	Global plan of action (1996)
GPS	Global positioning system
GRASS	Geographic resources analysis support system
GR/CIDS	Genetic resources communication, information and documentation system (1976)
GUI	Graphical user interface

GxE	Genotype by environment interaction
IARC	International agricultural research center
ICBN	International code of botanical nomenclature (1753)
ICNCP	International code of nomenclature for cultivated plants (Cultivated plant code) (1953)
IPT	Global Biodiversity Information Facility (GBIF) Integrated publishing toolkit (2009)
ITPGRFA	International treaty on plant genetic resources for food and agriculture (2001)
IU	International undertaking on plant genetic resources for food and agriculture (1983)
kNN	k nearest neighbor
KWIK	Keyword in context
LDA	Linear discriminant analysis
MATLAB	Matrix laboratory
MCPD	Multi-crop passport descriptors
NBP	Nordic biometry project
NOBIS	Nordic biometry system
N-PLS	Multilinear partial least squares (multiway PLS)
MAA	Most appropriate accession
MOS	Most original accession
PCA	Principal component analysis
PCR	Principal component regression
PGR	Plant genetic resources
PGRFA	Plant genetic resources for food and agriculture
PhD	Doctor of philosophy (*philosophiae doctor*)
PHP	Hypertext preprocessor
PLS-DA	Partial least squares discriminant analysis
PPVFR	Protection of plant varieties and farmers' rights (India, 2001)
RF	Random forests
RWA	Russian wheat aphid (*Diuraphis noxia* Kurdjumov)
SCUD	Swedish cultivated plants database [*Svensk kulturväxt database*, SKUD]
SDM	Species distribution model
SESTO	Seedstore management system
SIMCA	Soft independent modeling of class analogies (1976)
SINGER	System-wide information network for genetic resources of the CGIAR
sMTA	Standard material transfer agreement (2004/2006)
SOTW	Report on the state of the world's plant genetic resources for food and agriculture (1997)
SOTW2	Second report on the state of the world's plant genetic resources for food and agriculture (2010)
SVM	Support vector machines
TAXIR	Taxonomic information retrieval (1969)
TRIPS	Trade related aspects of intellectual rights (1994)
Ug99	Stem rust, race Ug99 (*Puccinia graminis* Pers.)
URL	Uniform resource locator
WorldClim	Global climate layers

Institutes and networks:

AEGIS	A European Genebank Integrated System (2009)
ASA	American Society of Agronomy, Madison, Wisconsin (1907)
ASARECA	Association for Strengthening Agricultural Research in Eastern and Central Africa (1993)
AWCC	Australian Winter Cereal Collection (Tamworth) (1967)
BioCASE	Biological Collection Access Service for Europe (2001-2004)
BioGeomancer	BioGeomancer project (2005-2007)
Bioversity	Bioversity International (2006) [IBPGR, 1974; IPGRI, 1991]
CAC	Plant Genetic Resources in Central Asia and the Caucasus (1998)
CATIE	Centro Agrónomico Tropical de Investigación y Enseñanza, Turrialba, Costa Rica (1973)
CBD	Convention on Biological Diversity (1992/1993)
CENARGEN	Centro Nacional de Pesquisa de Recursos Genéticos e Biotecnologia, Brasilia, Brazil
CGN	Centre for Genetic Resources, the Netherlands, Wageningen, Netherlands (1985)
CGRFA	FAO Commission on Genetic Resources for Food and Agriculture (1983)
CIAT	*Centro Internacional de Agricultura Tropical*, Cali, Colombia (1967)
CIMMYT	*Centro Internacional de Mejoram del maíz y del trigo*, Mexico City, Mexico (1966)
CRC	Carlsberg Research Center (1976) [Carlsberg Laboratory, 1875]
CSSA	Crop Science Society of America (1955)
CSIRO	Commonwealth Scientific and Industrial Research Organization, Australia (1926)
EAPGREN	Eastern Africa Plant Genetic Resources Network (2003) [1997]
ECPGR	European Cooperative Programme for Plant Genetic Resources (1980)
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária, Brasilia D. F., Brazil (1975)
EPGRIS	Establishment of a European Plant Genetic Resources Information Infra-Structure (2000-2003)
EUCARPIA	European Association for Plant Breeding Research (1956)
FAO	Food and Agriculture Organization of the United Nations, Rome, Italy (1948)
GCP	Generation Challenge Programme (2003-2013)
GBIF	Global Biodiversity Information Facility (2001)
GRDC	Grains Research and Development Corporation (1990)
IBP	International Biological Program (1964-1974)
IBPGR	International Board for Plant Genetic Resources (1974-1991)
ICARDA	International Centre for Agricultural Research in the Dry Areas, Aleppo, Syria (1976)
INRA	Institut National de la Recherche Agronomique (1946)
IPCC	Intergovernmental Panel on Climate Change (1988)
IPGRI	International Plant Genetic Resources Institute (1991-2006)
IPK	*Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung* [Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben] (1992) [1943]
[IPK]	*Kaiser-Wilhelm-Institut für Kulturpflanzenforschung* [Kaiser Wilhelm Institute of Crop Plant Research] (1943-1992)
ISTA	International Seed Testing Association (1924)
JIC	John Innes Centre [John Innes Horticultural Institution, 1910]
KU	University of Copenhagen [*Københavns Universitet*] (1479)
KU LIFE	University of Copenhagen, Faculty for Life Sciences (2007) [KVL, 1856]
KVL	The Royal Veterinary and Agricultural University [*Den Kongelige Veterinær- og Landbohøjskole*] (1856) [*Veterinærskolen*, 1773]
MTT	Agrifood Research Finland, Jokioinen, Finland (2001) [1898]

NGB	Nordic Gene Bank, Alnarp, Sweden (1979-2007) [reorganized as NordGen in 2007]
NGH	Nordic Genebank for Farm Animals (1984-2007) [reorganized as NordGen in 2007]
NSFP	Nordic Council for Forest Reproductive Material (1970-2007) [reorganized as NordGen]
NordGen	Nordic Genetic Resource Center (2007) [NGB, NGH, and NSFP was reorganized as NordGen in 2007]
PGR Forum	European Crop Wild Relative Diversity Assessment and Conservation Forum (2003-2005)
SEEDNet	South East European Development Network on Plant Genetic Resources (2004)
SGSV	Svalbard Global Seed Vault (2008)
SINGER	System-wide Information Network for Genetic Resources (1997)
SIS	Svalbard International Seedbank (planned in 1989 but not realized)
SNSK	*Store Norske Spitzbergen Kullkompani* [Great Norwegian Spitzbergen Coal Mining Company] (1916)
SSSA	Soil Science Society of America, Madison, Wisconsin (1936)
SYNTHESYS	Synthesis of Systematic Resources [EU] (2004-2009)
TDWG	Biodiversity Information Standards (1985) [previously: Taxonomic Databases Working Group]
UN	United Nations (1945)
UPOV	International Union for the Protection of New Varieties of Plants (French: *Union internationale pour la protection des obtentions végétales*), Genève, Switzerland (1961)
USDA	United States Department of Agriculture (1862)
USDA ARS	USDA Agricultural Research Service (1953)
USDA NPGS	USDA ARS National Plant Germplasm System (1990)
USDA GRIN	USDA ARS NPGS Genetic Resources Information Network
[VIR]	Bureau of Applied Botany (1894-1916), under the 'Scientific Committee'
[VIR]	Department of Applied Botany and Plant Breeding (1916-1924)
[VIR]	All-Union Institute of Applied Botany and New Crops (1924-1930)
VIR	The All-Union Institute of Plant Industry (1930-1968)
VIR	N.I. Vavilov All-Union Institute of Plant Industry (1968-1992)
VIR	N.I. Vavilov Research Institute of Plant Industry (1992), St. Petersburg
VASKhNIL	V.I. Lenin All-Union Academy of Agricultural Sciences of USSR (1929-1992)
ВАСХНИЛ	Всесоюзная академия сельскохозяйственных наук имени В. И. Ленина (Russian)
UNEP	United Nations Environment Program (1972)
WTP	World Trade Organization (1994)

International Agricultural Research Centers (IARC):

In chronological order

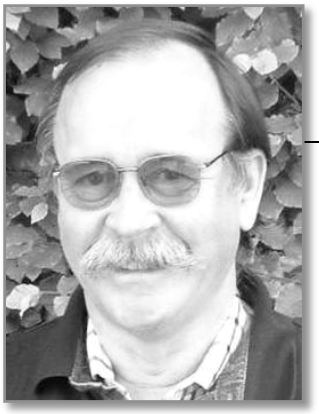
IRRI	International Rice Research Institute (**1960**) Los Baños, Philippines
CIMMYT	International Maize and Wheat Improvement Centre (**1966**) El Batán, Mexico
CIAT	Centre Internacional de Agricultura Tropical (**1967**) Cali, Colombia
IITA	International Institute of Tropical Agriculture (**1967**) Ibadan, Nigeria
CGIAR	*Consultative Group on International Agricultural Research (**1971**) [Umbrella for the IARCs]*
AfricaRice	Africa Rice Center (2009) Cotonou, Benin [WARDA, **1970**-2009]
CIP	International Potato Centre (**1970**) Lima, Peru
ICRISAT	International Crops Research Institute for the Semi-Arid Tropics (**1972**) Hyderabad, India
Bioversity	Bioversity International (2006) Rome, Italy [IBPGR, **1974**-1991; IPGRI, 1991-2006]
ILRI	International Livestock Research Institute (1994) Nairobi, Kenya [ILRAD, ILCA (**1974**-1994)]
IFPRI	International Food Policy Research Institute (**1974**) Washington DC, USA
ICARDA	International Centre for Agricultural Research in the Dry Areas (**1975**) Aleppo, Syria
WorldFish	World Fish Center (2002) Penang, Malaysia [ICLARM, **1977**-2000]
ICRAF	World Agroforestry Centre (**1977**) Nairobi, Kenya [new name in 2002, but same acronym]
IWMI	International Water Management Institute (**1996**) Battarmulla, Sri Lanka [ILMI, 1985-1996]
CIFOR	Center for International Forestry Research (**1993**) Bogor, Indonesia

Not active, reorganized and renamed IARCs:

IBPGR	International Board for Plant Genetic Resources (1974-1991) Roma, Italy [renamed to IPGRI in 1991; renamed to Bioversity International in 2006]
ICRAF	International Centre for Research in Agroforestry (1977-2002) Nairobi, Kenya [renamed to World Agroforestry Center in 2002, keeping the same acronym]
IPGRI	International Plant Genetic Resources Institute (1991-2006) Roma, Italy [reorganized together with INIBAP in 1994, and renamed to Bioversity International in 2006] [IBPGR, 1974]
INIBAP	International Network for the Improvement of Banana and Plantain (1984-1994) Montpellier, France [reorganized together with IPGRI in 1994, which was renamed to Bioversity in 2006]
ISNAR	International Service for National Agricultural Research (1980-2004) The Hague, Netherlands [closed and transferred to IFPRI in 2004]
IIMI	International Irrigation Management Institute (1984-1996) Battarmulla, Sri Lanka [renamed to IWMI in 1996]
ICLARM	International Center for Living Aquatic Resources Management (1977-2000) Philippines [renamed to WorldFish in 2000, and relocated to Malaysia]
ILRAD	International Laboratory for Research on Animal Diseases (1973-1994) Kenya [reorganized together with ILCA as ILRI in 1994]
ILCA	International Livestock Centre for Africa (1974-1994) Addis Ababa, Ethiopia [reorganized together with ILRAD as ILRI in 1994]
WARDA	West Africa Rice Development Association (1970-2009) Bouaké, Côte d'Ivoire [renamed to AfricaRice in 2009, and relocated to Benin]

Timeline of selected events:

1845 - The great Famine in Ireland (potato genetic diversity failure)
1905 - International Agricultural Institute (Rome, predecessor to FAO)
1945 - United Nations (UN)
1945 - Food and Agriculture Organization of the United Nations (FAO)
1961 - The International Union for the Protection of New Varieties of Plants (UPOV)
1961 - The Green Revolution in India (dwarf wheat and semi-dwarf rice)
1972 - United Nations Stockholm Conference on the Human Environment
1974 - International Board for Plant Genetic Resources (IBPGR)
1979 - Nordic Gene Bank (NGB)
1983 - Commission on Genetic Resources for Food and Agriculture (CGRFA)
1983 - International Undertaking on Plant Genetic Resources (IU)
1992 - Convention on Biological Diversity (CBD)
1994 - Agreement on Trade-Related Aspects of Intellectual Property (TRIPS)
1996 - Leipzig Global Plan of Action (GPA)
1996 - State of the World's Plant Genetic Resources for Food And Agriculture (SOTW)
2001 - International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA)
2004 - ITPGRFA enters into force 29 June 2004
2008 - Nordic Genetic Resource Center (NordGen)
2008 - Svalbard Global Seed Vault (SGSV)
2010 - The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture (SOTW2)
2010 - CBD Nagoya protocol on Access and benefit sharing (ABS)



Appendix 2: Sir Bent Skovmand

Bent Skovmand (1945-2007) was responsible for starting my PhD thesis; and one of the talented young scientists that Borlaug called upon to join the team in Mexico (Dworkin, 2009). Bent was an expert on stem rust by training (Skovmand, 1973, 1976). He was born on 25 January 1945 in Denmark and was in 1966 accepted for entry as a student to the Minnesota Agricultural Student Trainee Program including classes at the University of Minnesota, Saint Paul campus. In 1969 he enrolled at the Department of Plant Pathology. He completed his master on stem rust in 1973, and his PhD on the same topic in 1976. During the studies Bent made an impression on Emeritus Professor Elvin C. Stakman (1885-1979) who arranged a meeting between Bent and Norman Borlaug. After Bent completed the PhD he was immediately recruited to the wheat programme at CIMMYT in Mexico. In 1979 Bent was put in charge of the Triticale breeding programme. Between 1984 and 1988 Bent joined a wheat program in Turkey on secondment from CIMMYT. When he returned to CIMMYT in 1988 he became the head of the CIMMYT wheat genebank. One of his major achievements at the wheat genebank was the development of the International Wheat Information System (IWIS), made available for free on two CD-ROM discs.

In 2003 Bent first received the Frank N. Meyer from the Crop Society of America, and later the same year he was knighted by Queen Margrethe II of Denmark receiving the Knight's Cross of the Order of Danebrog.

In the beginning of 2004, we had the pleasure to welcome Bent as the new director at the Nordic Gene Bank in Alnarp, Sweden. Among many valuable contributions as the new director at NGB, Bent contributed to the development of the plans for the Svalbard Global Seed Vault. The Nordic Gene Bank safe duplication seed storage at Svalbard was already operative for 20 years when the plans for an international backup storage was opened again on the initiative from the IPGRI in Rome on behalf of the CGIAR, and in collaboration with NorAgric located at Ås outside Oslo. It was a sad loss when Bent died on 6 February 2007. Author Susan Dworkin wrote a book about Bent 'The Viking in the Wheat Field' (Dworkin, 2009) a book much recommend for learning more about plant genetic resources, wheat or of Sir Bent Skovmand. Later the same year as the book, Zeyen and Groth (2009) wrote a memorial for Bent on behalf of the American Phytopathological Society.

Bent Skovmand, Michael Mackay, and Kenneth Street initiated the first plans for my thesis project in the spring of 2004. In November 2006, Bent Skovmand announced the start-up of the PhD project leading to this thesis at the 2006 International Meeting of the Crop Science Society of America (CSSA) organized in Indianapolis, Indiana, USA (Endresen and Skovmand, 2006). The real start-up of the thesis research was in September 2007 in collaboration between the Nordic Genetic Resources Center (NordGen) and Bioversity International based in Rome.

Photo of Bent Skovmand by Lars Falk June 2006. NordGen Picture Archive image 003881.

Appendix 3: Bibliography of FIGS

This appendix provides an overview with the present status of published FIGS studies. This list includes the scientific papers published in scientific journals, and the abstracts presented at scientific conferences and symposia.

Mackay, M.C. (1986), Utilizing wheat genetic resources in Australia, [Conference proceedings]

Mackay, M.C. (1990), *Strategic planning for effective evaluation of plant germplasm*, [Book chapter].

Mackay, M.C. (1995), *One core collection or many?* [Book chapter].

Street, K., and M. Mackay (2003-2004), *Revolutionizing Plant Genetic Resources Management, Trait Discovery & Utilization, A Project Concept Note*, [Booklet, limited circulation. The term: FIGS is coined here].

Skovmand, B., M.C. Mackay, and K. Street (2004), *A new approach to locating and utilizing oat genetic resources*, [Conference proceedings (1), 7th Int Oat Conf, Jokioinen, Finland].

Mackay, M.C., and K. Street (2004), *Focused identification of germplasm strategy – FIGS*, [Conference proceedings, 11th Wheat Breeders Assembly, Sept 2004, Canberra, ACT, Australia]

Street, K., E. De Pauw, J. Ryan, and M.C. Mackay (2004), *Focused Identification of Germplasm Strategy: Identifying Wheat Landraces for Salinity Screening in Eurasia*, [Conference proceedings, ASA CSSA SSSA Int Annual Meetings, Nov 2004, Seattle, WA, USA].

Konopka, J., I. Kosareva, M. Mackay, O. Mitrofanova, K. Street, P. Strelchenko, J. Valkoun, E. Zuev, M.F. Nawar (2005-2007), *FIGS - Focused Identification of Germplasm Strategy, Bread Wheat Landrace Database*, [web site, http://figstraitmine.com/]

Endresen, D.T.F., and B. Skovmand (2006), *Trait Mining in Gene Banks*, [Conference proceedings, ASA CSSA SSSA Int Annual Meetings, Nov 2006, Indianapolis, IN, USA].

Street, K., M. Mackay, E. Zuev, N. Kaul, M. El Bouhssini, J. Konopka, O. Mitrofanova (2008), *Swimming in the genepool - a rational approach to exploiting large genetic resource collections*, [Conference proceedings, 11th Int Wheat Gen Symp, Aug 2008, Sydney, Australia].

Berger, J., J.A. Palta, C. Ludwig, D. Shrestha, M.C. Mackay, K.A. Street, J. Konopka, S. Jenkins, K.N. Adhikari, H.C. Clarke, J.S. Sandhu, and H. Nayyar (2008), *Emerging opportunities for agriculture: investigating plant adaptation by characterizing germplasm collection habitats*, [Conference proceedings, 14th Agro Conf, Sep 2008, Adelaide, South Australia].

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**, published April 2009].

El Bouhssini, M., K. Street, A. Joubi, Z. Ibrahim, and F. Rihawi (2009), *Sources of wheat resistance to Sunn pest,* Eurygaster integriceps *Puton, in Syria*, [**GRACE**, published April 2009]

Endresen, D.T.F. (2010), *Predictive association between trait data and ecogeographic data for Nordic barley landraces*, [**Crop Science**, published Sept 2010].

El Bouhssini, M., K. Street, A. Amri, M. Mackay, F.C. Ogbonnaya, A. Omran, O. Abdallah, M. Baum, A. Dabbous, and F. Rihawi (2010), *Sources of resistance in bread wheat to Russian wheat aphid (*Diuraphis noxia*) in Syria identified using the Focused Identification of Germplasm Strategy (FIGS)*, [**Plant Breeding**, published Oct 2010].

Endresen, D.T.F., K. Street, M.C. Mackay, A. Bari, and E. De Pauw (2011), *Predictive association between biotic stress traits and ecogeographic data for wheat and barley landraces*. [Crop Science, accepted April 2011]

Appendix 4: Trait mining algorithms and MATLAB

MATLAB code and scripts used for the calculations of PAPER III and IV.
Available online at: http://code.google.com/p/trait-mining/, http://goo.gl/i52HL

- Step 1: Import dataset to MATLAB
- Step 2: Calibrate and tune trait-mining models
- Step 3: Predict trait scores in the test set
- MATLAB scripts

Step 1: Import dataset to MATLAB

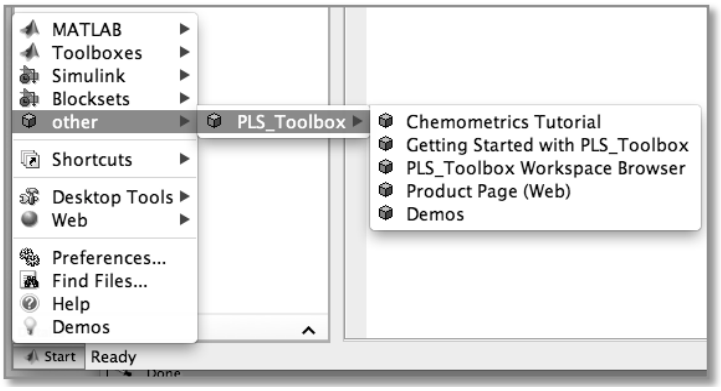
Source data:
USDA_Net_blotch_agroclim.xls (17.2 MB) [4651 records x 225 columns; 17 tables]
USDA_Stem_rust_agroclim.xls (16.6 MB) [10829 records x 142 columns]
Yemen_Ug99_agroclim.xlsx (3.4 MB) [4635 records x 133 columns]

Delete rows without coordinates and other rows without ecoclimate data.

USDA net blotch set: 2786 records/accessions (distinct sites not generated)
USDA stem rust set: 6889 records, 4932 accessions, 2013 collecting sites
Yemen Ug99 set: 4563 records/accessions, 1928 collecting sites
Training set (scores included): 825 records/accessions (20%)
Test set (scores hidden): 3738 records/accessions

MATLAB (R2010b)
http://www.mathworks.com/help/toolbox/stats/

PLS Toolbox 5.8 (R5.8.2)
http://software.eigenvector.com/toolbox/ (software)
http://wiki.eigenvector.com/index.php?title=Main_Page (documentation)
http://software.eigenvector.com/faq/ (faq)

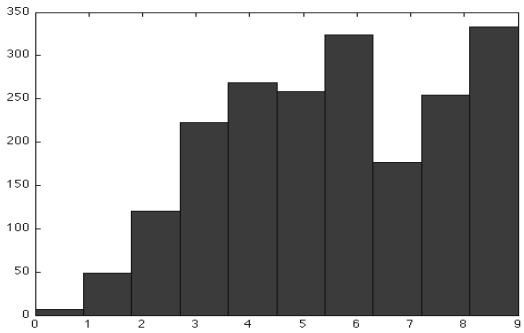


MATLAB -> Menu -> File -> Import Data… <sr_usda.txt> % Text columns before data columns

```
sr_usda = dataset(data(:, 1:115)); % build dataset, DSO, select data rows and columns
sr_usda.name = 'sr_usda';
sr_usda.author = 'Dag Endresen';
sr_usda.description = 'USDA Stem rust dataset'
sr_usda.title{1} = 'accessions';
sr_usda.label{1,1} = textdata(2:end,24);
```

```
sr_usda.labelname{1,1} = textdata(1,24); % sr3
sr_usda.label{1,2} = textdata(2:end,12); sr_usda.labelname{1,2} =textdata(1,12); % site_code
sr_usda.label{1,3} = textdata(2:end,1); sr_usda.labelname{1,3} = textdata(1,1); % accnumb
% … etc
sr_usda.class{1,1} = str2num(sr_usda.label{1,1}); sr_usda.classname{1,1} = 'sr3'; % sr3 (OK)
sr_usda.class{1,1} = data(:,117); sr_usda.classname{1,1} = 'sr3'; % sr3 (OK)
sr_usda.class{1,2} = textdata(2:end,12); sr_usda.classname{1,2} =textdata(1,12); % site_code
% … etc
sr_usda.title{2} = 'variables'; % second mode
% Column header names and type were prepared and imported to MATLAB
sr_usda.label{2,1} = sr_usda_columns(32:146); sr_usda.labelname{2,1} = 'variable'; % climate variable
sr_usda.label{2,2} = sr_usda_columns(32:146,2); sr_usda.labelname{2,2} = 'type'; % climate variable
type

% -- Test for distinct class values
unique(sr_usda.class{1,1}); % unique values first class
unique(sr_usda.label{1,1}); % unique values first label
hist(sr_usda.class{1,1}); % histogram
```




```
% Sample PI 212925 was identified as an outlier, and was removed
% -- Triticum aestivum ssp. aestivum; Country: India; Site: IND-5250
% -- Longitude: 73.66667; latitude: 17.91667; altitude: 1230 m
% -- Stem rust susceptibility, sr9: 9; sr3: 3; St Paul (MN), 1992
col = str2num(sr_usda.label{1,3}); % read acno from label
key = find(col(:,1) == 212925); % find PI 212925 (acno 212925)
sr_usda.data(key,:); % To display sample data
sr_usda(key,:).label; % To display sample labels
% sr_usda = delsamps(sr_usda, key, 1, 1); % Last flag (1) indicates a soft delete
sr_usda = delsamps(sr_usda, key, 1, 2); % Last flag (2) indicates a hard delete

% -- Replace values in an array
s = [1 1 1 0 0];
s(~s==1)=-1;  % replace values NOT (~) equal (==) to 1 with -1
s(s==0)=-1; % replace values equal (==) to 0 with -1

% -- Find missing values (NaN) in the X matrix
find(sum(isnan(X.data)') >=1); % display row keys with NaN
find(sum(isnan(X.data')') >=1); % display column key with NaN

% -- 3-way data cube (Results not reported here)
sr3w_wc(:, :, 1) = sr_usda(:, 61:72); % Mode 3, variable 1: Precipitation (WorldClim)
sr3w_wc(:, :, 2) = sr_usda(:, 73:84); % Mode 3, variable 2: Minimum temperature (WorldCim)
sr3w_wc(:, :, 3) = sr_usda(:, 85:96); % Mode 3, variable 3: Maximum temperature (WorldClim)
```

Step 2: Calibrate and tune the trait-mining models

```
% -- Select Unit
X = sr_usda; % USDA stem rust set, multiple observations per accession
% X = sr_usda_sites; % USDA stem rust set, distinct sites

% -- Select EcoClimate set
X = X(:, 1:48); X.name = cat(2,X.name,'_icarda4'); % (prec, tmax, tmin, pet) 48 columns
% X = X(:, 1:36); X.name = cat(2,X.name,'_icarda3'); % (prec, tmax, tmin) – 36 columns
% X = X(:, 49:84); X.name = cat(2,X.name,'_wc'); % (tmin, tmax, prec) 36 columns
% X = X(:, 85:103); X.name = cat(2,X.name,'_bio'); % Bioclim  - 19 columns

% -- Stratified subset (eventual)
% X = X(find(str2num(X.label{1,16}) == 11) ); X.name = cat(2,X.name,'_t_aestivum');
% X = X(find(str2num(X.label{1,16}) == 21) ); X.name = cat(2,X.name,'_t_durum');
% X = X(find(str2num(X.label{1,18}) == 1) ); X.name = cat(2,X.name,'_stpaul');
% X = X(find(str2num(X.label{1,18}) == 2) ); X.name = cat(2,X.name,'_rosemount');
% X = X(find(str2num(X.label{1,17}) == 1988) ); X.name = cat(2,X.name,'_1988');
% X = X(find(str2num(X.label{1,17}) == 1989) ); X.name = cat(2,X.name,'_1989');
% X = X(find(str2num(X.label{1,17}) == 1991) ); X.name = cat(2,X.name,'_1991');
% X = X(find(str2num(X.label{1,17}) == 1992) ); X.name = cat(2,X.name,'_1992');
% X = X(find(str2num(X.label{1,17}) == 1993) ); X.name = cat(2,X.name,'_1993');
% X = X(find(str2num(X.label{1,17}) == 1994) ); X.name = cat(2,X.name,'_1994');

% -- Split in random training set + test set
clear Xcal; clear Xtest; % Clean any previous training and test set from memory
s = floor(size(X,1)/3); r = randperm(size(X,1));  % split + random permutation
Xcal = X(r(s+1:end), :); Xcal.name = cat(2,X.name,'_cal'); % Training set
Xtest = X(r(1:s), :); Xtest.name = cat(2,X.name,'_test');  % Test set
dso_info(X); dso_info(Xcal);  dso_info(Xtest); % Displays information on the screen

% -- AUTOMATIC pre-study
tm_pre_study (Xcal, Xtest); % Fully automatic pre-study tuning tests

% -- DA (discriminant analysis) // included in the automatic pre-study
dso_info(Xcal); dso_info(Xtest); fprintf(1,'LDA classification:\n----------\n')
pclass=classify(Xtest.data,Xcal.data,Ycal,'linear'); pred=[Ytest',pclass]; pred2kappa(pred);
% // Diag-covar-matrix TYPE: 'linear' LDA, 'diaglinear' DA-DL, 'diagquadratic' DA-DQ

% -- kNN (k Nearest Neighbors) // included in the automatic pre-study
clear pclass; pclass = knn(Xcal, Xtest, 1);
dso_info(Xcal); dso_info(Xtest); fprintf(1,'kNN classification:\n----------\n');
clear pred; pred = [Ytest', pclass]; pred2kappa (pred);

% -- SIMCA // included in the automatic pre-study
% sr_simca (Xcal, Xtest, 7); % last input is the number of PCs
tm_simca_loop (Xcal, Xtest, 20); % last input is the number of PCs to loop

% -- PLS-DA // included in the automatic pre-study
% tm_plsda (Xcal, Xtest, 7); % last input is the number of LVs
tm_plsda_loop (Xcal, Xtest, 15); % last input is the number of LVs to loop

% -- PLS regression
% tm_pls (Xcal, Xtest, 15); % last input is the number of LVs
tm_pls_loop (Xcal, Xtest, 20); % last input is the number of LVs to loop
```

Step 3: Predict trait scores in the test set

- Using the model parameters identified from the 'pre-study' to the dataset
- Calibrate the model with the training set (known trait scores)
- Pre-processing: autoscale; Cross-validation: venetian blinds, 10 splits
  - With the so-called Venetian blinds schema and 10 splits, the 1st sample, the 11th sample, the 21th sample, ... is kept out in the first cross-validation set; then the 2nd, 12th, 22th, … is held out, and so forth.
- Apply the model to the samples with unknown trait scores (test set)

```
X = 0; % NO X for the Yemen predictions
Xcal = sr_yemen_826; % 826 samples from the Yemen Ug99 set for calibration
Xtest = sr_yemen_obs; % complete Yemen Ug99 set
Xcal = Xcal(:, 1:48); Xcal.name = cat(2,Xcal.name,'_icarda4'); % (prec, tmax, tmin, pet) 48 cols
Xtest = Xtest(:, 1:48); Xtest.name = cat(2,Xtest.name,'_icarda4'); % (prec, tmax, tmin, pet) 48 cols
Xcal.name = cat(2,Xcal.name,'_sr3');
Xcal.class{1,1} = Xcal.class{1,3}; Xcal.classname{1,1} = 'sr3';
Xcal.label{1,1} = Xcal.label{1,3}; Xcal.labelname{1,1} = 'sr3';

% -- Y response variable (actually using Xcal DSO first class for SIMCA)
Ycal = Xcal.class{1,1}; % class Y for SIMCA (one row)
Ytest = zeros(size(Xtest.class{1,1},2),1)'; % empty array with zeros, one column

% -- kNN (k Nearest Neighbour)
pclass = knn(Xcal, Xtest, 1);

% -- SIMCA
simca_options = simca('options'), simca_options.display = 'off';
simca_options.plots = ('final', 'none');
simca_options.preprocessing = {preprocess('default','autoscale')};
% simca_model = simca(Xcal, [], simca_options); % triggers the GUI
simca_model = simca(Xcal, [15], simca_options); % Highlighted is the number of PCs
simca_pred = simca(Xtest, simca_model); % options not needed for predictions
pclass = simca_pred.nclass'; % save predictions
% plotgui(simca_pred); % Opens a GUI to explore the SIMCA prediction model

% -- LDA (this classifier was not used for the reported prediction results)
pclass=classify(Xtest.data, Xcal.data, Ycal, 'linear'); % LDA

% -- PLS-DA (this classifier was not used for the reported prediction results)
Y_cal = class2logical(Ycal); % logical Y for PLSDA
Y_cal.class{1,1} = Xcal.class{1,1}; Y_cal.classname{1,1} = Xcal.classname{1,1};
Y_cal.label{1,1} = Xcal.label{1,1}; Y_cal.labelname{1,1} = Xcal.labelname{1,1};
Y_test = 0; % NO Y_test for the Yemen set
plsda_options = plsda('options'); plsda_options.plots = ('final', 'none');
plsda_options.preprocessing={preprocess('default','autoscale') preprocess('default','meancenter')}
plsda_model = plsda(Xcal, Y_cal, 9, plsda_options);
plsda_pred = plsda(Xtest, plsda_model) % no Y_test for the Yemen data
% plsda_pred = plsda(Xtest, Y_test, plsda_model, options_plsda) % validation if Y_test available
yprob = plsda_pred.detail.predprobability;
ymax = max(yprob')'; for i=1:(size(yprob,2)-1), ymax = cat(2, ymax, max(yprob')'); end;
yl = (yprob == ymax); % prob 2 logical (yl)
pclass = zeros(size(yprob,1),1); % empty array with zeros, one column
u = [1 2 3]; % [1 9]
% u = unique(Y_test.class{1,1}); % read class from the DSO
for i=1:(size(yprob,2)), pclass(find(yl(:, i) == 1)) = u(1,i); end;
```

MATLAB script:

The trait mining algorithms developed for this thesis were implemented as the following MATLAB scripts. Available online at: http://goo.gl/i52HL

- dso_info.m
- pred2kappa.m
- kappa_less.m
- tm_simca.m
- tm_simca_loop.m
- tm_plsda.m
- tm_pre_study.m

Script: dso_info.m

```
function dso_info(X)
% ** DSO_INFO **
% This function displays the name of the DSO and the first class.
%
% Syntax:    dso_info(X)
%
%   INPUT:     X - Dataset Object
%   OUTPUT:    Displays on the screen a summary of the Dataset Object
%   Example:   X = usda_t_aestivum(:, 97:115);
%
% Script dso_info by: Dag Endresen (dag.endresen@gmail.com), GPL2, 30 July 2010
% See also: KAPPA, CONFUSIONMAT, KNN, CLASSIFY, SIMCA, PLSDA
%

if isempty(X), error('Warning: X matrix is empty...'); end;
fprintf('-------------------------------------------------------\n');

% -- Display DSO name and size
fprintf(1, 'Dataset Object Name: \t%s \n', X.name);
[rows, cols] = size(X);
fprintf(1, 'DSO matrix size: \tRows = %0.0f, Columns = %0.0f \n', rows, cols);

% -- Display categories (first class)
fprintf('Categories (class 1): \t');
fprintf('%s: ', X.classname{1,1});
fprintf('%0.0f ', unique(X.class{1,1}));
fprintf('\n');

% -- Count samples
clear u; clear col; clear key; clear a;
u = unique(X.class{1,1});
col = X.class{1,1};
for i = 1:size(u,2),
    key = find(col(1,:) == u(1,i));
    a(i,1) = u(1,i);
    a(i,2) = size(key,2);
end;
disp(a');
return;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%% /dso_info %%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Script: kappa_less.m

Very minor modification of the MATLAB script **kappa.m** by Cardillo (2007).

Adding return parameters for the calculated Cohen's kappa and the observed agreement
"function [k, po] = kappa_less(varargin)"

Cardillo G. (2007) Cohen's kappa: compute the Cohen's kappa ratio on a 2x2 matrix.
Available online from: http://www.mathworks.com/matlabcentral/fileexchange/15365

Script: pred2kappa.m

```
function [k, po, k2, po2, pa, ppv, spec, sens, tp, fp, fn, tn] = pred2kappa(pred)
% ** PRED2KAPPA **
% This function computes and displays the Cohen's kappa coefficient
% To calculate the KAPPA coefficient we first need a Confusion Matrix
% Array of actual class against predicted class as a cross-tab
%
% Syntax: [k,po, k2,po2, pa,ppv, spec,sens, tp,fp,fn,tn] = pred2kappa(pred)
%
%   INPUT:
%     pred - array of actual, and predicted_class: [a p; a p; ...; a p]
%
%     From the classification models, the predicted class is extracted,
%     and combined as the second column together with the actual class.
%     This new array has the samples down as rows and two columns.
%
%   OUTPUT:
%     Displays on the screen a summary including the confusion matrix and
%     the KAPPA coefficient output from the script created by
%     Giuseppe Cardillo (giuseppe.cardillo-edta@poste.it).
%     Available online from:
%     http://www.mathworks.com/matlabcentral/fileexchange/15365
%     Cardillo G. (2007) Cohen's kappa: compute the Cohen's kappa ratio
%     on a 2x2 matrix.
%
%     Cohen's Kappa (k), Observed Agreement (po)
%
%     And for the matrix collapsed to a 2x2 confusion matrix:
%      Positive Agreement (pa), Positive Predictive Value (ppv),
%      Specificity (spec), Sensitivity (sens), True Positives (tp),
%      False Positives (fp), False Negatives (fn), True Negatives (tn)
%
%   Example:
%       pred=[2 1; 3 3; 1 1; 1 1; 1 3; 2 1; 2 2; 3 2];
%       pred2kappa(pred)
%
% Script pred2kappa by: Dag Endresen (dag.endresen@gmail.com), GPL2, 30 July 2010
% See also: KAPPA, CONFUSIONMAT, KNN, CLASSIFY, SIMCA, PLSDA, dso_info
%

if isempty(pred), error('Warning: PRED matrix is empty...'); end;
if isvector(pred), error('Warning: PRED must be a matrix not a vector'); end;

% Confusion matrix // error matrix
[C, order] = confusionmat (pred(:,2)', pred(:,1)');
disp('Confusion matrix:');
disp(C);
disp('Class agreement (per class):');
class_agreement = cat(1,order',(diag(C) ./ sum(C')')'); % ratio per class
disp(class_agreement);

% kappa(C,n) % 0=unweighted, 1=linear weighted, 2=quadratic, -1=all
[k1, po1] = kappa_less(C,0); % kappa_less based on Cardillo (2007)
[k, po] = kappa_less(C,1);
```

```
agreement_no = sum(diag(C)); % sum samples agree
agreement_po = sum(diag(C)) / sum(sum(C)); % ratio agree
fprintf('Observed agreement (Num)     = %0.0f samples \n',agreement_no);
fprintf('Observed agreement (PO)      = %0.3f\n',agreement_po);
fprintf('Cohen''s kappa, no wgt (k)    = %0.3f\n',k1);
fprintf('-- Linear Weighted --\n');
fprintf('Cohen''s kappa, weighted (k)  = %0.3f  <--- Kappa (weighted)\n',k);
fprintf('Observed agreement (po)      = %0.3f  <--- PO (weighted)\n',po);

% -----------------
% Collapse C to 2x2
% -----------------
if (size(C,1)==3),
    disp('-----------------------------------------------------------');
    disp('Confusion matrix collapsed from 3x3 to 2x2:');
    C2 = [C(1,1), sum(C(1,2:3));
        sum(C(2:3,1)), sum(sum(C(2:3,2:3)))];
    disp(C2);
elseif (size(C,1)==9),
    disp('-----------------------------------------------------------');
    disp('Confusion matrix collapsed from 9x9 to 2x2:');
    C2 = [sum(sum(C(1:3,1:3))), sum(sum(C(1:3,4:9)));
        sum(sum(C(4:9,1:3))), sum(sum(C(4:9,4:9)))];
    disp(C2);
elseif (size(C,1)==2),
    C2 = C;
    disp(C2);
elseif (size(C,1)>=5),
    disp('-----------------------------------------------------------');
    disp('---------- WARNING: missing categories ----------------------');
    disp('- Make sure that categories 1, 2, 3 are NOT missing!!!');
    disp('- Calculations will ONLY be valid with three categories!!!');
    disp('-----------------------------------------------------------');
    disp('Confusion matrix collapsed from NxN to 2x2:');
    C2 = [sum(sum(C(1:3,1:3))), sum(sum(C(1:3,4:end)));
        sum(sum(C(4:end,1:3))), sum(sum(C(4:end,4:end)))];
    disp(C2);
else
    disp('--------- WARNING: Problem with the Confusion matrix -------');
    C2 = [1 1; 1 1]; % Dummy C to avoid the script crashing below
end;

if ~isempty(C2)
    tp = C2(1,1); % True Positives
    tn = C2(2,2); % True Negatives
    fp = C2(1,2); % False Positives
    fn = C2(2,1); % False Negatives
    po2 = sum(diag(C2)) / sum(sum(C2)); % PO
    pa = 2*tp / ( 2*tp + fp + fn ); % PA
    ppv = tp / ( tp + fp ); % PPV
    sens = tp / ( tp + fn ); % Sensitivity
    spec = tn / ( fp + tn ); % Specificity
    [k2, po2_k] = kappa_less(C2, 1);
    % -- DISPLAY indicators
    fprintf('Cohen''s kappa       (k, 2x2) = %0.3f\n',k2);
    fprintf('Observed agreement (PO, 2x2) = %0.3f\n',po2);
    fprintf('Observed pos. agr. (PA, 2x2) = %0.3f  <--- PA \n',pa);
    fprintf('Positive pred. val (PPV,2x2) = %0.3f  <--- PPV \n',ppv);
    fprintf('Specificity       (Spec,2x2) = %0.3f  <--- Specificity \n',spec);
    fprintf('Sensitivity       (Sens,2x2) = %0.3f \n',sens);
    fprintf('-----------------------------------------------------------\n');
    fprintf('Cohen''s kappa, weighted (k)  = %7.3f  <--- Kappa (weighted)\n',k);
    fprintf('Observed agreement wgt (po)  = %7.3f  <--- PO (weighted)\n',po);
    fprintf('True Positives     (TP, 2x2) = %7.0f  <--- TP \n',tp);
    fprintf('True Negatives     (TN, 2x2) = %7.0f  <--- TN \n',tn);
    fprintf('\n');
else
    tp = NaN; tn = NaN; fp = NaN; fn = NaN;
    pa = NaN; ppv = NaN; sens = NaN; spec = NaN; po2 = NaN; k2 = NaN;
```

```
end;

return;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%% /pred2kappa %%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Script: tm_simca.m

```
function [k,po,k2,po2,pa,ppv,spec,sens,tp,fp,fn,tn] = tm_simca(Xcal, Xtest, pc)
% ** SIMCA model for trait mining (FIGS) **
% This function will run a SIMCA classification
%
% Syntax: [k,po,k2,po2,pa,ppv,spec,sens,tp,fp,fn,tn] = tm_simca(Xcal, Xtest, pc)
%
%   INPUT:
%     Xcal - Calibration set (DSO)
%     Xtest - Test set (DSO)
%     pc - number of principal components for the calibration
%        - can be a scalar or a vector with the PC for each class
%
%   OUTPUT: Displays on the screen a summary of the classification
%   Example: tm_simca(Xcal, Xtest, [4 4]);
%
% Script by: Dag Endresen (dag.endresen@gmail.com), GPL2, 3 August 2010
% See also: dso_info, pred2kappa, KAPPA, CONFUSIONMAT, KNN, CLASSIFY, SIMCA, PLSDA
%

if isempty(Xcal), error('Warning: Xcal matrix is empty...'); end;
if isempty(Xtest), error('Warning: Xtest matrix is empty...'); end;
if isempty(pc), error('Warning: pc scalar/vector is empty...'); end;

fprintf('-----------------------------------------------\n');
fprintf(1,'-------- SIMCA classification (PC %d) ---------\n', pc);
fprintf('-----------------------------------------------\n\n');

% -- Create response variable Y
Ycal = Xcal.class{1,1}; % class Y for SIMCA
Ytest = Xtest.class{1,1}; % class Y for SIMCA

% -- SIMCA OPTIONS
options = simca('options');
options.display = 'off'; % on/off
options.plots = 'final'; % final/none
options.preprocessing = { preprocess('default','autoscale') };

% -- SIMCA MODEL
simca_model = simca(Xcal, [pc], options);

% -- SIMCA PREDICTION (apply model)
simca_pred = simca(Xtest, simca_model);

% -- DISPLAY RESULTS
dso_info(Xcal); dso_info(Xtest);
fprintf(1,'SIMCA classification (PC %d):\n----------------------------\n', pc);
pclass = simca_pred.nclass'; pred = [Ytest', pclass];
[k, po, k2, po2, pa, ppv, spec, sens, tp, fp, fn, tn] = pred2kappa (pred); % Kappa
fprintf(1,'/SIMCA classification (PC %d)\n----------------------------\n', pc);

return;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%% /tm_simca %%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Script: tm_simca_loop.m

```
function pc = tm_simca_loop(Xcal, Xtest, pcs)
% ** SIMCA model for trait mining, FIGS **
% This function will run a SIMCA classification
%
% Syntax: pc = tm_simca_loop(Xcal, Xtest, pcs)
%
%   INPUT:
%     Xcal - Calibration set (DSO)
%     Xtest - Test set (DSO)
%     pcs - number of principal components for the last SIMCA model (loops)
%
%   OUTPUT: Displays on the screen a summary of the classification
%   Example:  tm_simca_loop(Xcal, Xtest, 20);
%
% Script by: Dag Endresen (dag.endresen@gmail.com), GPL2, 3 August 2010
% See also: dso_info, pred2kappa, KAPPA, CONFUSIONMAT, KNN, CLASSIFY, SIMCA, PLSDA
%

if isempty(Xcal), error('Warning: Xcal matrix is empty...'); end;
if isempty(Xtest), error('Warning: Xtest matrix is empty...'); end;
if isempty(pcs), error('Warning: pcs scalar is empty...'); end;

fprintf('\n');
fprintf('-----------------------------------------------\n');
fprintf(1,'-------- SIMCA classification (PC 1 to PC %d) ---------\n', pcs);
fprintf('-----------------------------------------------\n');

% -- DISPLAY RESULTS
dso_info(Xcal); dso_info(Xtest);
fprintf(1,'SIMCA classification (PC 1 to PC %d):\n------------------------\n', pcs);
% -- Loop PCs for finding SIMCA model complexity
n=0; a = zeros(pcs, 12);
for i = 1:1:pcs
    n = n+1;
    a(n, 1) = i;
    [a(n,2),a(n,3),a(n,4),a(n,5),a(n,6),a(n,7),a(n,8),a(n,9),a(n,10),a(n,11),a(n,12),
a(n,13)] = tm_simca (Xcal, Xtest, i);
end;
fprintf('PC  kappa  po      k(2x2)  po(2)   pa     pvv    spec   sens    ');
fprintf('tp    fp    fn   tn \n');
for n = 1:pcs
    fprintf('%2.0f %6.3f %6.3f %7.3f %7.3f %7.3f %6.3f %6.3f %6.3f %5.0f %5.0f %5.0f
%5.0f \n', a(n,1), a(n,2), a(n,3), a(n,4), a(n,5), a(n,6), a(n,7), a(n,8), a(n,9),
a(n,10), a(n,11), a(n,12), a(n,13) );
end
fprintf('\n');
bb = zeros(1,3); % initialize
% b is an array with the key(s) for the highest indicator values
% b(1) selects the first of the highest indicator values (simplest model)
% bb collects the "best" models (#pc) for PA, PVV, and Specificity
% median(bb) picks the middle/median #PC from PA, PVV, Specificity

b = find(a(:,2) == max(a(:,2))); % find key(s) for the highest Kappa
fprintf('\t Max Kappa: %6.4f   at PC: %2.0f \n', max(a(:,2)), b(1) );
b = find(a(:,6) == max(a(:,6))); bb(1,1) = b(1);
fprintf('\t Max PA   : %6.4f   at PC: %2.0f  <--- PA \n', max(a(:,6)), b(1) );
b = find(a(:,7) == max(a(:,7))); bb(1,2) = b(1);
fprintf('\t Max PVV  : %6.4f   at PC: %2.0f  <--- PPV \n', max(a(:,7)), b(1) );
b = find(a(:,8) == max(a(:,8))); bb(1,3) = b(1);
fprintf('\t Max Spec : %6.4f   at PC: %2.0f  <--- Spec \n', max(a(:,8)), b(1) );
b = find(a(:,9) == max(a(:,9)));
fprintf('\t Max Sens : %6.4f   at PC: %2.0f \n', max(a(:,9)), b(1) );
fprintf('---------------------------------------\n');
pc = median(bb);
fprintf('\t   Suggested number of PCs: %2.0f \n', median(bb) );
% -- DISPLAY indicators
```

```
fprintf('Cohen''s Kappa (K)    for PC %0.0f : %7.3f   <--- Kappa \n', pc, a(pc,2) );
fprintf('Observed Agreem (PO) for PC %0.0f : %7.3f   <--- PO \n', pc, a(pc,3) );
fprintf('True Positives (TP)  for PC %0.0f : %7.0f   <--- TP \n', pc, a(pc,10) );
fprintf('True Negatives (TN)  for PC %0.0f : %7.0f   <--- TN \n', pc, a(pc,13) );
fprintf('\n\n');
fprintf('NOTE THAT THIS IS ONLY A VERY ROUGH ESTIMATION OF MODEL COMPLEXITY\n\n');

return;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%% /tm_simca_loop %%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Script: tm_plsda.m

```
function [k,po,k2,po2,pa,ppv,spec,sens,tp,fp,fn,tn] = tm_plsda(Xcal, Xtest, lv)
% ** PLS-DA model for trait mining (FIGS) **
% This function will run a PLS-DA classification
%
% Syntax: [k,po,k2,po2,pa,ppv,spec,sens,tp,fp,fn,tn] = tm_plsda(Xcal, Xtest, lv)
%
%   INPUT:
%     Xcal - Calibration set (DSO)
%     Xtest - Test set (DSO)
%     lv - number of latent variables for the calibration (scalar)
%
%   OUTPUT:   Displays on the screen a summary of the classification
%   Example:  tm_plsda(Xcal, Xtest, [4 4]);
%
% Script by: Dag Endresen (dag.endresen@gmail.com), GPL2, 3 August 2010
% See also: dso_info, tm_simca, pred2kappa, KNN, CLASSIFY, SIMCA, PLSDA
%

if isempty(Xcal), error('Warning: Xcal matrix is empty...'); end;
if isempty(Xtest), error('Warning: Xtest matrix is empty...'); end;
if isempty(lv), error('Warning: lv scalar/vector is empty...'); end;

fprintf('-----------------------------------------------\n');
fprintf(1,'-------- PLS-DA classification (LV %d) ---------\n', lv);
fprintf('-----------------------------------------------\n');

% -- Create response variable Y
Ycal = Xcal.class{1,1}; % class Y for SIMCA
Ytest = Xtest.class{1,1}; % class Y for SIMCA (and pred below)
Y_cal = class2logical(Ycal); % logical Y for PLSDA
Y_test = class2logical(Ytest); % logical Y for PLSDA

% -- Add class and label to the logical Y for PLSDA
Y_cal.class{1,1} = Xcal.class{1,1};
Y_cal.classname{1,1} = Xcal.classname{1,1};
Y_cal.label{1,1} = Xcal.label{1,1};
Y_cal.labelname{1,1} = Xcal.labelname{1,1};

Y_test.class{1,1} = Xtest.class{1,1};
Y_test.classname{1,1} = Xtest.classname{1,1};
Y_test.label{1,1} = Xtest.label{1,1};
Y_test.labelname{1,1} = Xtest.labelname{1,1};

% -- PLS-DA OPTIONS
options = plsda('options');
options.plots = 'none'; % 'final', 'none'
options.preprocessing = {preprocess('default','autoscale')
preprocess('default','meancenter')};

% -- PLS-DA MODEL
plsda_model = plsda(Xcal, Y_cal, lv, options);
```

```
% -- PLS-DA PREDICTION (apply model)
% plsda_pred = plsda(Xtest, plsda_model, options); % no Y_test // blind test set
plsda_pred = plsda(Xtest, Y_test, plsda_model, options); % validation

% -- DISPLAY RESULTS
dso_info(Xcal); dso_info(Xtest);
fprintf(1,'PLS-DA classification (LV %d):\n---------------------------\n', lv)

yprob = plsda_pred.detail.predprobability;
ymax = max(yprob')';
for i=1:(size(yprob,2)-1), ymax = cat(2, ymax, max(yprob')'); end;
yl = (yprob == ymax); % prob 2 logical (yl)
pclass = zeros(size(yprob,1),1); % empty array with zeros, one column
u = unique(Y_test.class{1,1}); % read class from the DSO
for i=1:(size(yprob,2)),  pclass(find(yl(:, i) == 1)) = u(1,i);  end;

pred = [Ytest', pclass]; % actual-c, predicted-c
[k, po, k2, po2, pa, ppv, spec, sens, tp, fp, fn, tn] = pred2kappa (pred); % Kappa

fprintf(1,'/PLS-DA classification (LV %d):\n---------------------------\n\n', lv)
return;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%% /tm_plsda %%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

Script: tm_pre_study.m

```
function tm_pre_study(Xcal, Xtest)
% ** PRE-study **
% This function calibrates different classification and discrimination
% models and displays the results on screen
%
% Syntax:     tm_pre_study(Xcal, Xtest)
%
%   INPUT:
%     Xcal  - Dataset Object, independent/predictor data for calibration
%     Xtest - Dataset Object, independent/predictor data for validation
%
%   OUTPUT:
%     Displays on the screen a summary of the pre-study classification and
%     discrimination tests (kNN, SIMCA, LDA, DA-DL, PLS-DA, and PLS)
%
%   Example:
%     X = data(:, 1:48); % climate data (prec, tmax, tmin, pet, ...)
%     X = data(:, 85:103); % BioClim climate data
%
%     X = X(find(str2num(X.label{1,16}) == 11) ); % subset from label 16
%     X = X(find(str2num(X.label{1,16}) == 21) ); % subset from label 16
%
%     -- SPLIT in two subsets, Xcal, Xtest (predictor dataset, as DSO)
%     -- SET CLASS, the first class of the DSOs is used as Y (response)
%
%     tm_pre_study (Xcal, Xtest); % performs KNN, SIMCA, LDA, DA-DL, PLS-DA, and PLS
%
% Script by: Dag Endresen (dag.endresen@gmail.com), GPL2, 20 August 2010
% See also: pred2kappa, KAPPA, CONFUSIONMAT, KNN, CLASSIFY, SIMCA, PLSDA
%

if isempty(Xcal), error('Warning: Xcal matrix is empty...'); end;
if isempty(Xtest), error('Warning: Xtest matrix is empty...'); end;

fprintf ('-----------------------------------------------------------\n');
fprintf ('---------------- NEW PRE-STUDY TEST ----------------------\n');
fprintf ('-----------------------------------------------------------\n\n');
```

```
Ycal = Xcal.class{1,1}; % Y response class for SIMCA
Ytest = Xtest.class{1,1}; % Y response class for SIMCA
dso_info(Xcal); % number of samples for the Training set
dso_info(Xtest); % number of samples for the Test set

fprintf('\n\n');
fprintf('-------------------------------------------------------------\n');
fprintf('---------------- PRESS SPACE TO CONTINUE ------------------\n');
fprintf('-------------------------------------------------------------\n\n\n'); pause;
fprintf (' ---------------------- \n');
fprintf (' -- RANDOM selection -- \n');
fprintf (' ---------------------- \n\n\n');
fprintf('A rough test to select random samples\n\n');
clear r; r = randperm(size(Xtest,1)); % random permutation
clear Xr; Xr = Xtest(r, :); % random order all test samples
clear pclass; pclass = Xr.class{1,1}';
fprintf(1,'RANDOM classification:\n-------------------\n')
Ytest = Xtest.class{1,1};
clear pred; pred = [Ytest', pclass]; pred2kappa (pred); clear Xr;

fprintf('\n\n');
fprintf('-------------------------------------------------------------\n');
fprintf('---------------- PRESS SPACE TO CONTINUE ----------------\n');
fprintf('-------------------------------------------------------------\n\n\n'); pause;
fprintf (' --------- \n');
fprintf (' -- kNN -- \n');
fprintf (' --------- \n\n\n');
clear pclass; pclass = knn(Xcal, Xtest, 1);
fprintf(1,'kNN classification:\n-------------------\n')
clear pred; pred = [Ytest', pclass]; pred2kappa (pred);

fprintf('\n\n');
fprintf('-------------------------------------------------------------\n');
fprintf('---------------- PRESS SPACE TO CONTINUE ------------------\n');
fprintf('-------------------------------------------------------------\n\n\n'); pause;
fprintf (' ----------- \n');
fprintf (' -- SIMCA -- \n');
fprintf (' ----------- \n\n\n');
fprintf(1,'SIMCA classification:\n----------------------\n');
% tm_simca (Xcal, Xtest, 7); % last input is the number of PCs
tm_simca_loop (Xcal, Xtest, 15); % last input is the number of loops

fprintf('\n\n');
fprintf('-------------------------------------------------------------\n');
fprintf('---------------- PRESS SPACE TO CONTINUE ------------------\n');
fprintf('-------------------------------------------------------------\n\n\n'); pause;
fprintf (' ------------ \n');
fprintf (' -- PLS-DA -- \n');
fprintf (' ------------ \n\n\n');
fprintf(1,'PLS-DA classification:\n----------------------\n');
tm_plsda (Xcal, Xtest, 7); % last input is the number of LVs
% tm_plsda_loop (Xcal, Xtest, 15); % last input is the number of loops

fprintf('\n\n');
fprintf('-------------------------------------------------------------\n');
fprintf('---------------- PRESS SPACE TO CONTINUE ------------------\n');
fprintf('-------------------------------------------------------------\n\n\n'); pause;
fprintf (' -------- \n');
fprintf (' -- DA -- \n');
fprintf (' -------- \n\n\n');
% -- LDA is LAST because it often makes the script to crash
% -- //Error using ==> classify at 245
% -- //The pooled covariance matrix of TRAINING must be positive definite.
% -- diag-covar-matrix TYPE: 'linear' LDA, 'diaglinear' DA-DL, 'diagquadratic' DA-DQ
fprintf(1,'LDA classification:\n--------------------\n');
pclass=classify(Xtest.data,Xcal.data,Ycal,'linear');
pred=[Ytest',pclass]; pred2kappa(pred); % LDA

fprintf('\n\n');
```

```
fprintf('-------------------------------------------------------------\n');
fprintf('---------------- PRESS SPACE TO CONTINUE ------------------\n');
fprintf('-------------------------------------------------------------\n\n\n'); pause;

fprintf('\n\n');
dso_info(Xcal); dso_info(Xtest);

% plotscores(plsda_model, plsda_pred);
% plotsloads(plsda_model, plsda_pred);
% ploteigen(plsda_model, plsda_pred);
% plotgui(plsda_model, plsda_pred);

fprintf('\n\n');
fprintf('-------------------------------------------------------------\n');
fprintf('---------------- PRE-STUDY TEST COMPLETED -----------------\n');
fprintf('-------------------------------------------------------------\n\n\n');

return;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%%%%%%%% /tm_pre_study %%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

The genebank collections worldwide provide the genetic diversity required for plant breeding activities. To access the useful traits, the plant breeders most often need to conduct large screening experiments to identify the accessions holding such desired traits. Focused Identification of Germplasm Strategy (FIGS) was proposed as a new approach to assist plant breeders and other genebank users looking for a target trait property. The principles of this new approach is based on finding a link between the target trait property and the ecoclimatic profile at the original collecting location for landraces and wild relatives of the cultivated plants. This thesis provides some of the first experimental evidence to support the FIGS concept. Efficient crop improvement programs and effective plant breeding activities are a high priority to ensure future food supply under the challenges of climate change and a rapidly growing world population.

About the author, Dag Terje Filip Endresen.

Dag graduated from the Norwegian University of Science and Technology (NTNU) as Civil Engineer (Master of Technology) in 1996 with a specilization in Chemistry and Molecular genetics. After the military service and one year working for IBM, Dag started in 1999 to work with documentation of plant genetic resources at the Nordic Gene Bank (NGB) and took over as head of the documentation department at the institute in 2002. During these years the genebank information system at genebank institute was reorganized as an online database. NGB joined the Global Biodiversity Information Facility (GBIF) as member organization in March 2004. Based on the successful contribution with establishing new internet-based data sharing in the European Genebank Documentation Group, Dag made the opportunity in 2005/2006 to work on secondment for the International Plant Genetic Resources Institute (IPGRI) based in Rome. IPGRI is a partner of the Consultative Group on International Agricultural Research (CGIAR) and was in 2007 reorganized as Bioversity International. NGB was in 2008 reorganized as the Nordic Genetic resources Center (NordGen). The PhD research project reported with this book was organized in close collaboration between Copenhagen University, NordGen and Bioversity International, and also a contribution to the further development of the Focused Identification of Germplasm Strategy (FIGS).

PhD thesis. Utilization of Plant Genetic Resources: A Lifeboat to the Gene Pool.
Dag Terje Filip Endresen, Nordic Genetic Resource Center (NordGen).
University of Copenhagen, Faculty for Life Sciences, Department of Agriculture and Ecology.
Academic supervisor: Dvora-Laiô Wulfsohn and Brian Grout.
Submitted: 9 February 2011. Dissertation: 31 March 2011.

© Dag Terje Filip Endresen, dag.endresen@gmail.com
PDF version available online at: http://goo.gl/pYa9x
MATLAB source code available at: http://goo.gl/i52HL
Printed at Media-Tryck, Lund University Press, April 2011.

ISBN: 978-91-628-8268-6