WHITE-PAPER: BARLEY GENOME PHYSICAL MAP AND SEQUENCE

A PLATFORM FOR SYSTEMATIC GENE ISOLATION IN BARLEY

Nils Stein, Lothar Altschmied, Patrick Schweizer, Ulrich Wobus, Andreas Graner
Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK), Corrensstraße 3, 06466 Gatersleben, Germany

Robbie Waugh, David Leader, Luke Ramsay, Bill Thomas
Scottish Crop Research Institute (SCRI), Invergowrie, Dundee, DD2 5DA, Scotland

Alan Schulman
MTT Agrifood Research Finland and University of Helsinki, P.O. Box 56 (Viikinkaari 4), FIN-00014 Helsinki, Finland

Klaus Mayer, Hans-Werner Mewes
GSF Forschungszentrum, Ingolstädter Landstraße 1, 85764 Neuherberg, Germany
TU-München, Wissenschaftszentrum Weihenstephan, Am Forum 1, 85354 Freising, Germany

THE RATIONALE AND OBJECTIVE

To meet the needs of a growing world population and to support the development of a far-sighted bio-based economy, European agriculture has to continue its efforts in developing more effective strategies for agricultural innovation including alternative uses and novel products from existing crop plants. In addition, a sustainable agriculture requires the breeding for genetically tailored cultivars to reduce the input of agrochemicals and to adapt crop plants to a changing environment. Here, we propose the generation of a physical map-based sequence of the barley genome providing a reference point and a milestone in plant genetics, which will pay a major contribution for the development of any method suitable to employ comparative genome analysis for the deep understanding of the molecular processes underlying quality, yield, and resistance of plants - far beyond the catalogue of genes.

Barley ranks number five in world crop production (harvested area) with the European Community producing over 60 % of the world harvest (FAOSTAT 2004). Thus, any investment into this crop species will entail a major impact on European agriculture. In addition, barley combines the least complex *Triticeae* crop genome (diploid) with a wealth of genetic and genomics resources developed throughout the past decade (i.e. 400,000 ESTs, Affymetrix chip, transcript map with >1000 genes, isolation of 30,000 gene-containing BAC clones). Its genome exhibits a high colinearity with other *Triticeae* species (e.g. wheat, rye, raygrass) with which it also shares numerous agronomic traits. Thus, barley takes a central position not only as a crop plant but also as a model species for genetic studies within the *Triticeae*.

Variation among individuals is a fundamental requirement for genetics and has been the driving force behind both step change (e.g. improved yield via dwarf varieties) and incremental (e.g. quality) improvements made since crops were first domesticated. Over the past 100 years, conventional barley breeding has resulted in tremendously improved yield and disease resistance and due to a continuous genetic enhancement in brewing quality; the European brewing industries rely on malt produced from the domestic barley crop. However, further progress in breeding will critically depend on the knowledge of genes and on the systematic exploitation of their allelic diversity. The combination of genetic and genomic approaches will allow the elucidation of the genetic basis of major agronomic traits and facilitate a fundamental understanding of the underlying metabolic and developmental processes.

In barley, hundreds of regions of the genome confer improvement of specific qualities of the crop (yield, quality, stress tolerance etc.). The underlying genes (and their alleles) need to be identified, isolated and studied to understand how they exert their effect and to gain insight into how they can be efficiently utilized to further improve crop performance. In the majority of cases the most appropriate route to identifying and isolating these genes is by their position in the genome, which

in contrast to small, fully sequenced genome model plants like Arabidopsis and rice is time consuming and cumbersome. Furthermore, the majority of traits with importance to large genome cereal improvement cannot be studied in the small genome model plant. The work we propose will fundamentally change the current position, making positional gene isolation in barley both rapid and achievable. We anticipate that many of the barley genes that are targets for isolation control phenotypes that are directly comparable to phenotypes in its more complex and commercially important relative wheat.

The complete sequence of the barley gene space or at least the genomic sequences of most genes will be required to address integrative approaches to understanding crop plant performance. In order to provide the fundamental information of the genome, the sequence is needed to enable any other technology for the functional and structural exploration of any biological system. So far, any comparative approach to understand the complex functionality of the cereal genomes is handicapped by the lack of detailed information on the dynamics and interactions of their genetic elements. Common and distinguished functional and regulatory mechanisms between species (model and crop plants) can only be realized if multiple genome sequences become available as was learned from the human genome sequencing project.

At present the costs and technical difficulties do prevent an immediate start of genome sequencing at a large scale. Therefore, we favour and outline a stepwise approach building on established resources and technologies in which work packages will provide outputs for both the academic and commercial communities at each step on their own and which still allow to study biological problems in a timely manner. During the first phase a genome-wide physical BAC contig map will be constructed and anchored to a high-density genetic map facilitating instant access to basically any gene in the barley genome (WP 1, WP 2). The full exploitation of this resource will depend on the availability of deeply phenotyped plant material (germplasm and populations) developed in WP3. In WP 4 (Sample Sequencing) and WP 5 (Bioinformatics) fundamental issues on the architecture of the barley genome and the evolution of *Triticeae* genomes will be investigated in defined projects. Eventually, in a second phase the barley genome will be completely sequenced. This long-term goal will depend on the further improvement of sequencing and associated technologies and will require collaboration within an extended world-wide consortium. However, since barley represents a European model crop (esp. for the Triticeae), we are convinced that Europe should take the lead in unlocking the potential of the barley genome.

THE APPROACH

Our final goal is the sequencing of the barley genome, which we will approach in a stepwise manner. The project is expected to extend over a period of a decade as a concerted and multinational effort.
The work packages we propose are

Phase 1

- the construction of a BAC-based physical contig map (WP 1),
- the generation of a high-resolution genetic map (WP 2) on which the physical map will be anchored,
- the development of germplasm resources, mapping populations and comprehensive phenotyping (WP 3),
- sample sequencing of targeted regions in the barley genome (WP 4)
- the building of a bioinformatics infrastructure (WP 5)

Phase 2

- complete genome sequencing (WP 6).

WORK PACKAGE 1 – ESTABLISHING A PHYSICAL MAP:
Output: Preliminary (phase I) minimal tiling path of the barley genome
Benefits: Rapid and cost-effective isolation of nearly any gene
Timeframe: 24 (36) months
Existing resources: BAC library of cv. Morex (6x genome coverage, 100 kbp average insert size), 30,000 gene-containing BAC clones

The success of physical map construction will require a large-insert BAC-library with sufficient genome coverage (10x) and an average insert size exceeding 150 kb. A physical map will be constructed by high-throughput state-of-the-art fingerprinting analysis of approx. 350,000 clones. To exploit the available information on the gene space 30,000 gene containing BACs, identified within NSF and GABI projects from the existing BAC library of cv. Morex will be included in the fingerprinting procedure. Over 20,000 genes, including the genetically mapped (see work package2), will be placed on the physical map through overgo probe hybridisation.

Alternatively, the construction of subgenomic BAC libraries for each of the 14 barley chromosome arms would allow to reduce the complexity of the genome and split the task into smaller work packages, distributable to several labs allowing for a less complex contig map assembly. This would be achieved by cloning of DNA obtained from flow-sorted chromosomes - a strategy that has been successfully followed in the case of wheat chromosome 3B within a consortium between France and Czech Republic. The limitation of the procedure is that the cytogenetic stocks are available only for a cultivar Betzes for which specific genomics resources are not available neither would it allow to take advantage of any existing data obtained in cv. Morex so far.

WORK PACKAGE 2 – GENETIC ANCHORING OF THE PHYSICAL MAP:
Output: High-density and high-resolution genetic map of barley with 10,000 genes. An improved and deconvoluted (phase II) minimal tiling path of the physical map will be determined.
Benefits: Efficient genetic targeting of traits; resource for comparative genetics to other *Triticeae* and cereal species; minimal tiling path of the physical map.
Timeframe: 36-72 months
Existing resources: Genetic maps of barley with over 4,000 gene based (EST derived) markers (published and unpublished data IPK, SCRI, Kyoto University).

None of the available maps comprises the necessary genetic resolution and the required marker density by itself. A high resolution mapping population including Morex as one of the parental genotypes will be generated either by increasing resolution and marker density of existing populations (Steptoe x Morex, Morex x Barke) and/or by design of an “intermated recombinant inbred (IRI)” population to ensure for a high number of crossing over events reducing the overall necessary population size. High-throughput marker saturation will build on the information of 3,000-4,000 experimentally and several thousands *in silicio* determined Single Nucleotide Polymorphisms (SNP). For direct linking of assembled contigs, BACend sequencing of representative clones (up to 50,000) will provide information for additional ~1,000 markers (10% of low copy/genic sequences expected) to be added to the genetic anchor map.

WORK PACKAGE 3 – DEVELOPING OF GERMPLASM AND POPULATIONS:
Output: populations for mapping and isolation of genes for major agronomic traits
Benefits: Phenotypic information required for the identification and isolation of the underlying genes
Timeframe: Throughout the project
Existing resources: Genetic Resources from ex situ collections, well-characterized cultivars, pedigree information

The value of a physical map for map-based isolation of genes depends on the availability of phenotypic information. To unlock the genetic basis of key agronomic traits with either special importance to barley or general importance to all *Triticeae* crop species, comprehensive genetic resources and mapping populations will be developed and evaluated by a concerted community effort. This will include populations for association genetics, QTL mapping populations and near isogenic lines. Methods streamlined according to the technological development for standardised phenotyping of populations and germplasm will be developed in context of dedicated projects within a network organised under i.e. EU FP7 as well as supported by national and industrial funding.

WORK PACKAGE 4 – SAMPLE SEQUENCING THE BARLEY GENOME:
Output: 30 Mbp skim sequence for estimating the gene space of the barley genome, 15 Mbp sequence for detailed analysis of Triticeae genome organization
Benefits: gene distribution in the barley genome; gene and promoter sequences in regions of prime interest; detailed information on colinearity of the barley and wheat genome
Timeframe: 24 - 36 months
Existing resources: gene-containing BAC clones

Pilot sequencing will be dedicated to determine central aspects of barley genome organization. Independent approaches (deletion/translocation mapping, gene-based marker screening of BAC library) revealed an uneven distribution of genes occupying potentially less than 10 % space of the whole barley genome, thus sequencing of only the "gene space" could provide detailed information for a majority of the barley genes. 200 gene-containing BACs of barley will be skim-sequenced to three times coverage for determining total gene content and gene structure. Furthermore, strategies for selective assessment of genic regions like shotgun sequencing of genomic libraries developed by methyl-filtration (coding sequences are hypomethylated and can be enriched for in genomic libraries) will be evaluated for their potential of selectively sequencing the barley "gene space".

In order to determine overall *Triticeae* genome colinearity at large scale, 100 BAC clones, which were genetically anchored to chromosome 3H, and representing loci evenly distributed over the whole chromosome will be completely sequenced. This effort runs in parallel and will be coordinated with similar efforts recently started for wheat chromosome 3B (Dr. C. Feuillet, INRA Clermont-Ferrand, France) and will lead to a much more profound and precise picture of gene content and genome evolution between homoeologous Triticeae genomes.

WORK PACKAGE 5 – BIOINFORMATICS:
Output: dedicated tools for annotation of the *Triticeae* genomes, data warehouse of *Triticeae* genome information
Benefits: refined genome annotation; capacity building
Timeframe: Throughout the whole project
Existing resources: experienced bioinformatics groups at GSF, IPK, and SCRI

Bioinformatics will allow for the interpretation of physical map assembly, sequence/genome annotation, and will support database curation and exploitation of publicly available genomics data information. This will rely on well-developed tools originating from various international genome physical mapping and sequencing projects. Furthermore the specific aspects of *Triticeae* sequence and genome organisation will require developing of new tools for annotation of gene structure as well as for the plethora of different repetitive DNA elements which will screw up many of the sequence assembly steps. The accumulated information will be connected at any possibility to public genomics resources available for other grass species, *i.e*. the complete sequence of the rice genome, a physical map of the maize genome and upcoming resources for the sorghum and wheat genomes within a data warehouse environment for *Triticeae* genomics, which will provide essential cereal crop information for application in research and applied breeding. Activities will take place in close conjunction with established international centres for genome annotation.

WORK PACKAGE 6 – COMPLETE GENOME SEQUENCING OF BARLEY:
Output: genomic sequence information of the majority of barley genes, structure of the barley gene space
Benefits: complete sequences of all genes and their promoters; physical map position of agronomic traits; design of improved genechips for expression analysis; advanced possibilities for analysis of gene expression data; improved assignment of proteomics and functional genomics data; complete candidate gene lists for any genomic region of interest
Timeframe: 7 – 10 years
Existing resources: 3 Mb of genomic sequence

The physical map will be the starting point for the elucidation of the genomic sequence. Although the large extent of non-informative sequence will remain a major handicap, ongoing efforts to sequence maize indicate that combined strategies will allow generating a complete sequence at reasonable costs. Based on the Minimal Tiling Path sequencing of up to 3 -5,000 gene containing and physically/genetically anchored clones, which sum up to 10% of a haploid barley genome equivalent (estimated costs today = 20 Mio €) will make the start, providing incremental information about the "gene space" of the barley genome. This will cover a significant portion of all barley genes and relying on novel techniques to scan clones for their content of functional genetic elements the sequencing will be continued in further increments till completing the whole barley genome.

COLLABORATIONS

Aiming at sequencing a large *Triticeae* genome such as barley will require international collaboration. The strategic importance of barley for northern European countries suggests a leading role of Europe for initiating the project and thus will form a trigger to tie together national and international activities under one umbrella. While IPK is prepared to coordinate the effort one important aim is to facilitate participation of both small and large groups in a dynamic way *i.e.* according to their financial support. This will strengthen the links within the research community and accelerate technology transfer and assist human capacity building.

COLLABORATIONS WITHIN EUROPE

This program initiated within an existing consortium and currently coordinated by IPK Gatersleben (Germany) gathers together institutions with strong and long lasting reputation in barley genome analysis combining all necessary expertise to start with the proposed activities. The already participating teams foster intensive interactions with the international *Triticeae* Research Community. The diversity of aspects opens numerous possibilities for interaction and participation of interested research teams providing all requirements of forming a barley genomics network entering into the upcoming call for national and international funding of research activities in Europe (i.e. EraNet-Plant Genomics, FP 7).

INTERNATIONAL COLLABORATIONS

Strong and longer lasting collaborations are well established in the barley genomics and within the *Triticeae* community although funding for international collaborations is difficult and underdeveloped. Resources for barley genome analysis are intensively shared between European teams and partners in USA, Australia and Japan. A project of the proposed dimension can become a full success only if resources of funding, labour, equipment and expertise are merged. Milestone projects will rely, build on or incorporate intermediate results from international activities e.g. like the local physical map construction of gene rich regions funded by the NSF, USA. For the bioinformatics support, Germany has established a leading position through its essential contributions to the sequencing of *A. thaliana* and its ongoing part in the maize genome project. Therefore, coordinating the data collection, analysis, and interpretation relying on an established proven infrastructure will help to ensure the success of the project.

FUNDING

The aim of the presented initiative is to underline the fundamental importance of the availability of a BAC-based physical map and subsequently the genomic sequence of barley for efficient future cereal crop improvement. This issue of international relevance has to be addressed by an international effort and funding. All participating partners are convinced that the proposed approach is necessary and timely. To achieve the final goal of a physical map-based barley genome sequence, national and international funds have to be raised to achieve a solid support. This will include dedicated research projects funded within the national genome programmes of the participating partner institutions and will further rely on European efforts such as EraNet-PG, bi- and multilateral plant genomics projects and EU framework programme FP 7. Funding agencies are asked to open opportunities for the application of such a multinational initiative in plant genomics.