Three steps to complete access of the barley genome sequence

A revised concept for sequencing a 5 Gigabase genome in three steps of progressively improved quality

The International Barley Sequencing Consortium (IBSC), spring 2010

Preamble: "The introduction and maturation of next-generation sequencing technology is transforming our vision of large crop sequencing. It is no longer a question of 'can a crop genome be sequenced?' but rather, 'when will all major crop genomes be sequenced?'

Concept overview

The International Barley Sequencing Consortium (IBSC) published a whitepaper in 2006 proposing a research strategy for generating a high-quality reference sequence of the barley genome. It built on a generally accepted concept for development of a high-quality, or "gold standard" reference genome sequence including the following consecutive tasks:

- Construction of a genetically anchored physical map of the genome
- Application of "next generation" sequencing technology for its use in barley genome sequencing
- Clone-by-clone sequencing of barley chromosomes along a minimal tiling path of partially overlapping BAC clones

This concept, which was approved by representatives of several successful genome-sequencing projects, has recently been challenged by the success of draft sequencing moderately-sized plant genomes using the "whole genome shotgun" sequencing approach. The IBSC, together with the International Wheat Genome Sequencing Consortium (IWGSC), hosted a workshop (i.e., the IWGSC-IBSC Workshop on Sequencing Technologies, 11-12 September 2008, Genoscope, Paris-Evry, France) attended by representatives of the genome centers and informatics experts to discuss the most appropriate approaches to generate high quality reference sequences of wheat and barley.

Since 2006 further advances in sequencing and assembly technology have been introduced and the genomic infrastructure for barley has tremendously improved (see below). These new advances and the improved barley genomics infrastructure have been the impetus for revisiting the originally agreed genome-sequencing approach. This revised vision statement reflects the changing landscape and places the goals and deliverables and their downstream benefits into a series of realistically achievable milestones and timelines towards the overall goal of a high quality (gold standard) barley reference genome sequence.

Here, a three-step strategy towards a high quality reference genome sequence is described. It delivers important intermediate results and resources at each individual step and these will provide unique opportunities for research and application in barley and related cereal species (e.g., wheat, rye and Triticeae grasses):

Step 1: SURVEY SEQUENCE

Deliverables: sequence tags of virtually every barley gene; partial gene index information for each barley chromosome arm. **Application**: gene cloning, functional genomics, genome-wide marker development

Step 2: PHYSICAL MAP ANCHORED TO WHOLE GENOME SHOTGUN SEQUENCE

Deliverables: Draft genome sequence of barley consisting of thousands of small sequence contigs/scaffolds. Usefulness is strongly dependent on the depth of integration with the barley physical and genetic map; access to assembled sequence information of almost every barley gene as well as their flanking and local sequence contexts; knowledge of the entire gene complement. **Application**: Functional genomics / systems biology, genome wide survey for genetic diversity by resequencing, genome scan, genomic selection

Step 3: HIGH-QUALITY MAP-BASED REFERENCE GENOME SEQUENCE

Deliverable: High-quality sequence assemblies or "scaffolds" for the barley genome with minimum gaps. **Application**: complete genome re-sequencing; discovery of copy number variation and impact on trait variation in barley; access to non-genic regulatory regions; understanding epigenetic targets of the barley genome; ultimate resource for trait analysis; crop optimization by all means of modern crop improvement.

Current status of barley genome sequencing

Light touch coordination of IBSC partner research has assembled the essential tools and resources required for this revised barley genome sequencing strategy. This was achieved via combining the outputs from both nationally and internationally funded projects. The following list of resources is available or will be available by May 2010.

Physical map:	fingerprint assembly of >550,000 BAC clones at finalizing stage, >14x haploid genome coverage
Genetic anchoring:	1,424 BOPA1 and BOPA2 SNPs and 1500 genetically mapped EST amplicons anchored directly to BACs distributed throughout all seven barley linkage groups. In addition, over 10,000 genes (mostly without genetic map position) have been anchored to individual BAC addresses (and most to chromosome arms) by array hybridization to BAC pools (and chromosome arm preps). These are available for anchoring via conserved synteny with model grass genomes.
BAC end sequencing:	paired-end sequences of 300,000 BAC clones (50% of all clones entered into fingerprint assembly) are available (these sequence data alone represents ~8% of the barley genome)
BAC clone sequencing:	3500 BACs sequenced to Phase I; contained genes allow chromosomal anchoring for 2/3 of all sequenced BACs, provide anchoring templates for BAC end sequence data (these sequence data represents ~7% of the barley genome)
Whole genome survey:	1 – 10% of the barley genome produced as short reads (35 – 100 nt), analysis of genome composition and annotation

Sorted chromosome survey:	~1.5 x coverage shotgun sequence of every barley chromosome arm; ~70% of all barley genes were tagged by short sequence signatures; allows to assign these genes into chromosomal bins; allows anchoring of barley BACs to chromosomal bins
Expressed genes (EST):	complete or partial coding sequence available for 75% of all barley genes via sequencing of expressed sequence tags

This summary clearly documents that IBSC-coordinated research has effectively completed step 1 of the revised strategy. Thus, survey information at high resolution is available for the entire barley genome, providing novel und unique opportunities for research and application in the area of gene and trait isolation as well as marker development. Integration of the accumulated information lays the foundation for beginning Step 2 and provides the basis for later implementation of Step 3 of the barley sequencing concept.

What is required to proceed to the next steps

The datasets that are required to move towards Step 2 (whole genome shotgun sequencing) and Step 3 (map-based clone-by-clone sequencing) of the above outlined barley sequencing strategy will be generated by several approaches. In each case increased capacity and sequencing cost reduction provided by "next generation sequencing" technology will be exploited.

For whole genome shotgun (WGS) sequencing, data obtained from different sequencing platforms will be combined: high genome coverage (>50-fold haploid genome coverage) will be provided by short-read, paired-end and mate-pair sequencing (i.e. Illumina GA II or SOLiD 4) from progressively sized, sheared genomic DNA (0.5kb, 3kb, 8 kb, 20kb). This will be extended (an additional 10-fold genome coverage) by longer paired-end reads (e.g., Roche/454 Titanium) from several sizes of sheared genomic DNA (8kb, 20kb). The use of paired-end sequencing of progressively sized DNA fractions will allow assembly of gene-containing regions. It will also allow assembly of moderately repetitive regions of the genome or at least provide a scaffold of framework of genic segments separated by un-assembled repetitive DNA. Both shotgun datasets will be validated by the available 570,000 paired BAC-ends sequenced by the traditional Sanger method, and the Phase 1 assembled 454 BAC sequences. Together these will allow validation of the assemblies and ultimately provide the ability to develop scaffold sequence contigs over distances of between 100 and 150 kb. This, as well as the sequenced genes, will tightly anchor the entire sequencing dataset to the physical and genetic maps (which comprise >10,000 gene/BAC address relationships).

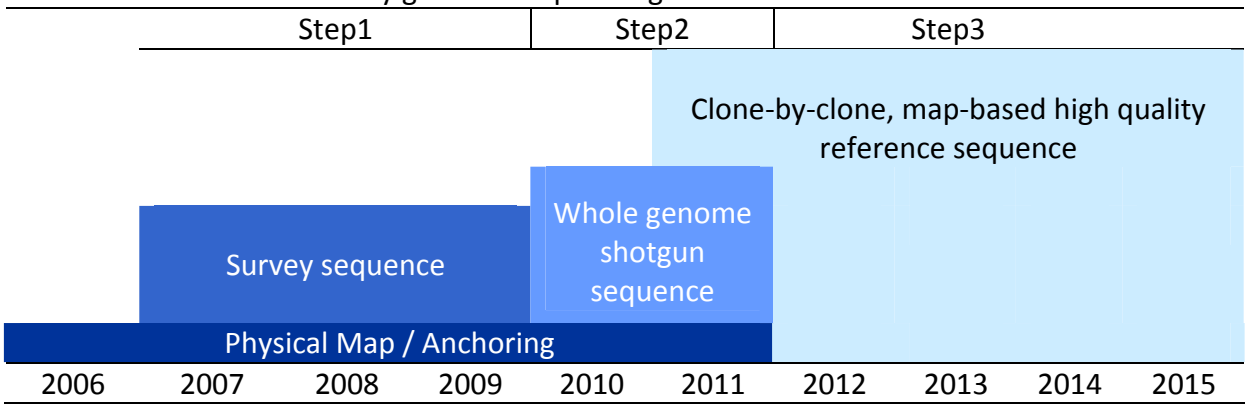
To achieve a more or less contiguous high quality reference genome sequence requires overlapping BAC clones derived from the minimal tiling path of the barley physical map to be sequenced. Based on the expected final quality of the physical map, between 5000 and 8000 BAC clones will comprise about 1000 to 1500 contigs per barley chromosome. Protocols for sequencing barley BAC clones on the available NGS platforms have been established. This approach will take advantage of sequence information generated during Steps 1 and 2, which will assist and improve the overall sequence assembly. Using todays technologies, this final stage will require significantly more labor and manual data curation and verification,

and will thus be substantially more time consuming. However we recognize that further advances in genome sequencing technologies and assembly algorithms may significantly simplify this task.

Timescale and expected costs for step-wise sequencing the barley genome

Work on Step 1 of the barley genome sequencing scenario started in 2006; this step was effectively completed by end of 2009 (Table 1). Today's technical advances provide the opportunity, within the time-frame of one year, to produce all sequence information required for accomplishing Step 2 – given that sufficient funding is available. The final stage, Step 3, towards a high quality sequence of the barley genome could be realized within the next five years, if it includes the necessary work of sequence finalization and gap closure. In the absence of any significant technological advances (alluded to above), this task will require a concerted international funding framework.

Table 1: Timeframe of barley genome sequencing



Step 1 of the barley sequencing project was reached by multinational funding totaling *c.* €10 million (Table 2). This figure includes the cost of EST sequencing, BAC library construction, development of high density gene maps, physical mapping and survey sequencing. The next step will require extensive shotgun sequencing. The cost of this task depends largely on what sequencing platforms are utilized and if sequencing is subcontracted to industry or performed within the academic research environment. A mixed model is likely, with the sequencing costs alone of Step 2 reaching approximately €2 million (table 2). At the current state, the mere sequencing costs for reaching Step 3 would be in the range of 2 million EURO per chromosome thus in total approximately 15 million Euro are required in Step 3 to reach to a high quality reference genome sequence.

Table 2: Overview of estimated sequencing costs (excluding personnel)

Step 1		
	High density genetic maps	2 million € (estimated)
	EST sequencing	2 million € (estimated)
	Physical mapping	3 million €
	Survey sequencing	3 million €
Total		10 million €
Step 2		
	80 x paired-end /mate pair sequencing Illumina GA II	0.5 million €
	10 x paired end 454 Titanium	1.5 million €
Step 3		
	Minimal Tiling path sequencing	2 million € / chromosome= 15 million €
TOTAL Step 2 & 3		17 million €