International Wheat Genome Sequencing Consortium

IWGSC – IBSC

INTERNATIONAL BARLEY SEQUENCING CONSORTIUM

Sequencing Technologies Workshop

Genoscope, National Sequencing Center
Evry, France

September 11-12 2008

Developing strategic roadmaps for sequencing the wheat and barley genomes

Sessions

- **Wheat and barley genomes: the challenge**
- **Experience from other genome sequencing projects**
- **Current wheat and barley sequencing projects**
- **How useful will be the new sequencing technologies?**
- **Bioinformatics**
- **Funding Strategies**

454 SEQUENCING

Applied Biosystems

illumina

INRA

NSF

IPK GATERSLEBEN

Bundesministerium für Bildung und Forschung

IWGSC – IBSC
Sequencing Technologies Workshop

International Wheat Genome Sequencing Consortium

INTERNATIONAL BARLEY SEQUENCING CONSORTIUM

Genoscope National Sequencing Center
Evry, France

September 11 - 12 2008

Workshop Report

Kellye Eversole, John Fellers, Catherine Feuillet, Gary Muehlbauer, Jane Rogers, Nils Stein

On September 11-12, 2008, approximately sixty researchers from around the world gathered at the Centre National de Séquençage, Genoscope, in Evry (France) for a workshop held under the auspices of the International wheat (IWGSC) and barley (IBSC) genome sequencing consortia to discuss sequencing technologies and strategic roadmaps for sequencing the wheat and barley genomes.

The first day started with introductions by Kellye Eversole (Eversole Associates), executive director of the International Wheat Genome Sequencing Consortia (www.wheatgenome.org), Nils Stein (IPK) Chair of the International Barley Sequencing Consortium (www.barleygenome.org), and Patrick Wincker of Genoscope. Jane Rogers, BBSRC, began with an overview of the workshop goals that included developing a white paper the international community would be able to use to advocate for funding the sequencing of the two genomes. She stressed the importance of conducting immediate pilot projects to enhance understanding of the genomes and pointed out the need for the engagement of bioinformatics capacities to reach the ultimate goals.

Catherine Feuillet, INRA (Clermont-Ferrand) and European co-chair of the IWGSC, began session one (Genome structure of wheat and barley) with an overview of the wheat and barley physical mapping and sequencing projects to date. The IWGSC, founded in 2005, has 40 coordinating committee members and more than 150 general members. It is governed by six co-chairs from five countries and one executive director (K. Eversole). The IBSC is chaired by Nils Stein (co-chair: Gary Muehlbauer, University of Minnesota), and comprises seven other members from six different countries. C. Feuillet also described the current progress of the wheat genome physical mapping project with results of the two most advanced projects (3B, 3AS). She described the chromosome-based strategy employed by the IWGSC to establish the physical maps of the 21 bread wheat chromosomes (cv. Chinese Spring) and provided details about the chromosome sorting technique, optimized by Jaroslav Dolezel, Institute of Experimental Botany (Czech Republic), that permits the isolation of each wheat chromosome/chromosome arms to produce specific BAC libraries. In her lab (INRA Clermont-Ferrand), C. Feuillet reported that 57,000 bacterial artificial chromosome (BAC) clones of a 3B-specific BAC library had been fingerprinted and assembled into 1036 contigs, with an average clone depth of 10-fold and which accounted for 82% of the chromosome. These were anchored to the genetic maps with 1443 genetic markers and a minimum tiling path (MTP) of 7500 BACs has been defined for further studies. A second fingerprinting effort and new MTP design is underway to reach 20x coverage in view of sequencing the chromosome. The chromosome 3A physical map is being developed by Kansas State University, under the direction of Bikram Gill. C. Feuillet reported that 44,500 BAC clones were fingerprinted representing a 12X coverage of the 3AS chromosome. These were assembled into 1677 contigs that were estimated to represent 269 megabases (Mb), approximately 75 percent of the chromosome.

A picture of gene distribution and content had been developed from some of the early BAC end sequencing efforts on 3B. C. Feuillet reported that from 20’000 BAC end sequences, it was estimated that the wheat genome comprises 86% repeat DNA of which 65% are LTR elements. The sequence has a GC content of 46.5% and an estimated 36,000 genes in the B genome. It was stated also that in contrast to previous suggestions, only 40% of the genes appear to be in gene-rich islands and that most genes are scattered throughout the genome. C. Feuillet reported that this finding was supported by the results of J. Bennetzen and K. Devos’ NSF funded project to sequence 200 random BACs from a Chinese Spring BAC library. In 4-5X shotgun sequence coverage of random clones, 28% of the BACs were without a recognizable gene sequence and 35% had only one gene present.

The second session (Experience from other genome projects) highlighted efforts and experiences from the sequencing of various animal and plant genomes. Jane Rogers discussed the lessons learned from the human and porcine genome sequencing projects. The human genome project was initiated in 1996 for a genome of an estimated size of 3 Gb with 40% of repeats. It was sequenced with 50 million Sanger-based reads using a strategy based on the development of a physical map with cosmids and BACs and a sequencing effort split among different countries at the international level. She compared the advantages and disadvantages of the Clone by Clone Strategy vs. the Whole Genome Shotgun (WGS) approach that was developed in 2000 by

Celera. The assembly of the WGS profited from the BAC by BAC approach. She pointed out that new resources (e.g. additional BAC libraries with different enzymes) had to be generated along the way and that for such large projects it is important to develop resources on demand as the project progresses. The sequencing of the MTP was done in 2 phases: (Phase 1) 4x Shotgun to provide a working draft for the whole genome with 400'000 contigs sequenced and using the physical map information to order the sequences and (Phase 2) additional sequencing to increase the quality and the coverage with additional shotgun and directed sequencing. The complete sequence released in 2003 was 2.85 Gb in which 99% of the euchromatin was sequenced to an accuracy of one error per 100,000 bases. J. Rogers stressed that the high quality of the sequence was extremely useful for further applications and that manual annotation is essential in this process. Finally, a genome browser was established to make the sequence accessible to the international community.

J. Rogers presented the porcine genome project that started in 2005 and was based on an exceptionally good physical map with 172 contigs that had been positioned using a high resolution RH map and alignment of BAC end sequences to the human genome. A combined approach of BAC by BAC (3-4 x) and WGS plus fosmid end sequences was chosen for sequencing, to maximize the value that could be achieved with the limited funding. This allowed the assembly of chromosome sequences (from assembled draft BAC sequences) that correspond to contig sequences ordered along the chromosomes. The sequencing is being carried out by the Wellcome Trust Sanger Institute and is expected to be completed in 2009.

André Eggen (INRA, Jouy en Josas) described the bovine genome sequencing project that started in 2001. He mentioned that the initial steps were the development of a white paper which was not very ambitious and induced a recurrent problem of funding in the following years as funding administrators countered that the goals of the white paper had been achieved. The project was supported strongly by the breeding industry. Several BAC libraries were produced from different breeds and a physical map was generated based on 15x coverage of a BAC library prepared from a single Hereford cow that was fingerprinted and assembled at the BC Cancer Agency. The sequencing strategy was a hybrid of BAC by BAC, WGS, and BAC/fosmids/plasmid ends sequencing. Drafts of the 3Gb "Dominette" Hereford breed sequence were released in 2004 (3x) and 2005 (6x), but were not useful since mapping data were not used to guide the assembly and there was no mechanism in place to allow the users to report misassembly problems to the sequencing center. Thus, a number of problems persisted in the assembly. A. Eggen pointed out the importance of establishing early in the project a mechanism to utilize feedback from the sequence users for improving the sequence assemblies and ensuring that the sequencing centers and users have similar goals thereby guaranteeing the production of a useful sequence. The first useful sequence, released in 2006, resulted from the incorporation of BAC and BES sequencing data. As soon as the reference sequence (6-12x) was produced, a SNP project was launched and several breeds were resequenced using reduced representation libraries for fast SNP discovery. To date 54,000 SNPs have been placed on the bovine genome.

Dan Rohksar, U.S. Department of Energy Joint Genome Institute (DOE-JGI), discussed sequencing efforts on soybean and sorghum crops, and the model plant *A. thaliana*. The genome sizes are 1.1 Gb, 750 Mb, and 350 Mb, respectively. In soybean, the genome was sequenced using Sanger reads with an average read length of 700 bp. The genome coverage was 8X and assembled into 973.3 Mb. All of the contigs were reconstructed into the 20 chromosomes, half of the loci were gap free, and 98% of the expressed sequence tags (ESTs) were placed on the contigs. Since soybean is an ancient tetraploid (10 MYA), there was a concern whether the sequences would be too similar and would produce a poor assembly. In practice, it was observed that sequence differences as low as 5% were sufficient to separate the homoeologs and no homologs were lost in the assembly. The sorghum genome is estimated to contain 26-28,000 genes and *E. coli*-based methyl filtration covered one level of genes, but missed another. In the sequence assembly, the pericentric portion of the chromosome was found to be the place of expansion and location of the majority of repeat DNA. Alignment of the rice and sorghum genomes revealed that genes were collinear up to the centromeres. Of the genes identified 4,000 appeared to be unique to sorghum and 50% of these were verified to be true genes. The intron sizes appeared to be similar in rice and sorghum and, of major significance, 98% of the introns were in the same position. With the development of the Roche 454 high throughput pyrosequencing platform, significant cost savings are attainable. D. Rohksar stated that the cost for Sanger based sequencing is now $1 US per 1 Kb and, for example, the cost of sequencing the maize genome in a 3 year project was estimated to be about $25m. Roche 454 titanium sequencing now costs $20 per Mb. To determine whether 454 sequences alone could be used for genome sequencing, the JGI had sequenced Arabidopsis *thaliana* (cv. *Cape Verde Island)* and generated10-fold genome coverage. Half of the assembled contigs were 26 Kb or larger and 97.2% of the sequence was found in the reference. The error rate was 1 in 14,400 bp.

Patrick Wincker, Genoscope, presented the efforts to sequence the grape genome (Pinot Noir). A whole genome shotgun (WGS) strategy with Sanger sequencing was chosen to sequence a homozygote variety to reduce the problems encountered during the physical mapping of the highly heterozygote commercial variety. A 12x sequence was obtained to fit the requirements for a high quality sequence from the users. Annotation of the genome sequence was facilitated by deep sequencing of cDNAs on Roche 454 FLX and Illumina sequencing platforms. 30,300 protein-coding genes were identified and, overall, genes cover 46.3% of the genome. Some of the most interesting discoveries were traces of ancient hexaploidization of the genome and the expansion of terpene synthesis gene families in relation with their importance in determining wine quality. Eighty-nine gene families were found to be associated with this pathway.

Pat Schnable, Iowa State University, talked about the maize genome sequencing project led by R. Wilson (Washington University, St. Louis). A first BAC library was made and fingerprinted to establish a physical map of B73 that assembled in 721 contigs, of which 421 were anchored to genetic maps. A number of contigs were mis-assembled and the importance of developing new assembly algorithms to better manage repeat-rich genomes was stressed. The $32 million project was not sufficient to

generate a fully sequenced genome using Sanger capillary sequencing methodology. Thus, 19'000 seed BACs were sequenced (16125 BAC sequences produced) with the goal of determining "the order and orientation" of the ~50'000 genes in maize. The BAC by BAC sequencing entailed generating 4-6x clone depth in shotgun reads and sequence improvement using directed reads. BAC end sequences (800'000) and fosmid end sequences were produced also. The BACs covered 93% of the genome and 50,000 genes were found. A high number (1%) of nearly identical paralogs (NIPs, 98% sequence identity)) were observed with many gene fragments. New algorithms were developed to break spurious alignments between the NIPs and detect paramorphims. An essential component of the project was to make the sequence accessible to the community as rapidly as possible, so that they could provide input to the project and improve the data, as well as have a platform where users could ask for the sequencing of their favorite BAC contig. A second library made from the cultivar Mo17 was shotgun sequenced and aligned to the reference B73 sequence. A visualization tool that highlights the confidence to which the assembly is secure is under development and is essential for the end users. P. Schnable emphasized the importance of having substantial user involvement in genome sequencing projects and recommended (i) setting up a community annotation working group early on in the project; (ii) developing tools to assess and display the quality of the sequence released, if the strategy is to give rapid access to the community; (iii) identifying or develop tools that display varying levels of certainty, and (iii) exploring improved assembly strategies.

The third session focused on "Current wheat and barley pilot projects". N. Stein presented work carried out on the Illumina Solexa platform. His group, in collaboration with the lab of D. Ware at CSHL, had analyzed 600 Mb of 35 nt reads to search for the frequency of discrete 20mer sequences within the barley genome with the aim of establishing a mathematically defined repeat (MDR) index (following the approach designed by D. Ware for maize). They found that 99% of the discrete 20mers occurred only 1 to 10 times and the remaining 1% of the discrete 20mer represent 20% of the barley genome. Comparison with finished BAC sequences indicated that statistical repeat annotation using the MDR index was, on one hand, very efficient and congruent to hand annotation of BAC sequences. On the other hand, it demonstrated that comprehensive knowledge has already been accumulated and developed about the shape of the repetitive DNA landscape of the barley genomes. Further comparisons with wheat showed that MDR of barley will not be useful for wheat and that MDRs are species specific. N. Stein's group has also developed a pilot project for sequencing BAC clones with- Roche 454 GS20 and GSFLX technologies. Using bar coded BACs and pooling strategies (50 BACs / pool), he obtained 20x coverage of clone sequences and compared the assemblies with finished sequences generated using Sanger sequencing technology. His results show that 454 GSFLX technology can be used to assemble whole BAC sequences. In more than 10 % of the cases full assemblies could be obtained, in many cases, the assembly resulted in 5-10 contigs per clone insert. Thus the sequencing strategy can deliver better than Phase-I draft sequences in the initial attempt and therefore can play a vital role in a sequencing strategy for complex genomes. In N. Stein's laboratory, the assembler MIRA performed better than Newbler (the Roche 454 assembler) but emphasized also the need to establish efficient advanced assembly algorithms for improved handling of next generation sequencing

data that will be obtained for the barley and wheat genome projects. N. Stein has investigated, in collaboration with the group of J. Doelezel, IEB Olomouc, using flow-sorted chromosomes to generate genome sequence for barley and has compared GSFLX sequence data of 1H with the rice genome. A bias towards genes on orthologous rice chromosome 5 and 1 indicated the potential of this approach. Finally, paired reads of up to 300,000 BAC clones (ERAPG BARCODE) and 25,000 FLcDNA clone sequences (K. Sato, T Matsumoto, Japan) are under production for the barley genome project.

P. Wincker, Genoscope, reported on a wheat BAC sequencing pilot project. Twenty-four tilled BACs, with an average insert size of 120 Mb, were sequenced to a depth of 20X using Roche 454 GSFLX technology. The clone sequences were assembled into 10-40 contigs per BAC, which is higher than had been obtained with other organisms at Genoscope. Thus, the GSFLX data alone would not generate complete high quality clone assemblies, , and Genoscope concluded that they would need to be complemented with single Sanger reads. This experience contrasted with that of N. Stein's barley BACs. Moreover, the hypothesis suggested by P. Winker that the differences might be due to different features present in the two genomes was questioned by several groups who have annotated BACs from both species. N. Stein proposed that the data should be re-assembled with MIRA and compared with the Newbler assemblies.

John Fellers, USDA-ARS Manhattan, Kansas, reported on a new technique for methyl filtration that could be used to reduce the representation of repetitive DNA in preparation for sequencing. Previous work with WGS sequencing in *Ae. tauschi*i showed that the 91% of the genome appeared to be repetitive. Using E. coli based methyl filtration and Cot reassociation kinetics, the amount of repetitive DNA was reduced to 74% and 31%, respectively. The new technique used methyl sensitive, six base cutting restriction enzymes. *Aat*II and *Pst*I reduced the amount of repetitive DNA to 16.2% and 19.1% respectively. It was suggested that this would be useful with the Roche 454 or Solexa platforms.

Session four (New Sequencing Technologies) consisted of presentations from four companies that are developing the Next Generation sequencing technology and equipment. These were, Roche 454 Life Sciences, Illumina, Helicos Biosciences, and Applied Biosystems. Lei Du from Roche 454 discussed the latest developments in the chemistry and sequencing instrument that will now support the generation of 300-500 base reads with the Titanium sequencing kits. Paired end (PE) reads are now systematically proposed for *de novo* sequencing projects with 3kb libraries routinely made and 16-20 kb libraries in development. A kit has been developed that permits tagging of up to 12 multiplexed samples and a 96 MIDs kit is under development (users can define their own MIDS as well). The strategy adopted for genomes with sizes up to 500 Mb is WGS and PE sequencing, while for more than 800 Mb, BAC by BAC sequencing is recommended using MTP and pooled BAC strategies combined with Sanger BAC end sequences. Pilot projects on potato and Salmon were reported. One of the bottlenecks in large genome projects currently is the throughput in library production. Roche is working on this aspect as a priority.

Sean Humphray from Illumina presented the developments that have been made to upgrade the GA1 sequencing platform. The GA2 can now support generation of 10-15 Gb data per run with 50-75nt reads. PE reads can be obtained with 2- 5kb fragment libraries and kits to support sequencing from larger fragment libraries (~10 kb) are in development. Protocols are available now for transcript and regulome analysis and 12-plex barcode tags are under development for multiplexing samples. Finally, the automation of the sampling process is also a priority for the next months.

Avak Kahvejian from Helicos, the most recent company to have released a new sequencing platform, presented their single molecule sequencing technology that can generate reads of 25 bp and produce 10Gb of sequence in 8 days. A recent publication in *Science* reported the use of the Helicos technology to sequence a viral genome. Although there is no need for amplification and ligation and the workflow is simpler than for Roche and Illumina, it is clear that Helicos is not ready to be used for *de novo* sequencing of large genomes at the moment.

A similar conclusion was reached by Marco Pirotta from Applied Biosystems who presented the SOLiD technology. SOLiD is a technology with ultra high throughput, typically generating 20 Gb/run in 25-50 nt paired reads (9 days) and employing a 2-base color coding protocol that increased the base call accuracy. This has proven value for resequencing since the base calling strategy ensures extremely high quality sequence and it is very useful for whole transcriptome analysis (including small RNAs, and exon-intron splicing analysis)

Session five (Bioinformatics) focused on presentations and discussions of the bioinformatic tools that are being developed to support the assembly and analysis of data from new sequencing technologies. Richa Agarwala, National Center of Biotechnological Information (NCBI) - National Institute for Health (NIH) USA, presented the resources the center has developed for mapping as well as for storage and retrieval of the new generation sequences. She first discussed the resources that her group has developed for high resolution radiation hybrid (RH) mapping that have been applied successfully in numerous animal genome sequencing projects and can assist greatly in sequence assembly and contig ordering. She explained the rationale and needs behind RH panel construction, marker selection, genotyping, and map construction (software packages) and illustrated this with examples from the cat and horse genome projects. She then presented the Window masker tool that has been developed to mask repetitive sequences in genomes for which no repeat library is available. Finally, she showed the new online submission system (SRA metadata) that has been developed for the submission of 454 and Illumina reads.

David Edwards, University of Queensland Australia, presented a comparison of the performances of the GSFLX, Illumina, and SOLiD platforms. In his experience, SOLiD has proved to be a highly effective platform for re-sequencing (with demonstrated ability to identify errors in Sanger assemblies) but that Illumina is better suited to *de novo* sequencing. He emphasized that read depth seemed more important than the read lengths, provided that it is possible to span repeats with paired end reads. He presented

the software SASSY that he has used to assemble the short reads from the SOLiD and Illumina platforms. The results from a pilot project to compare the sequence generated from pooled BACs using SOLiD and Illumina technologies had shown that the SOLiD sequence did not indicate an even distribution and had more mismatches when compared to a reference sequence. Although the Illumina sequence was more evenly distributed and had fewer errors, a higher level of variation between runs was observed. Base calls after 30 bases also tended to have a higher incidence of error. The SASSY program also was not able to resolve completely the true order of sequence contigs without additional ordering information. He pointed out that both platforms generate very large amounts of data that has to be taken into account.

Mike Bevan (JIC) presented briefly the VELVET assembler developed by E. Birney (EBI) as well as the objectives of a UK project to re-sequence BACs from Ph1 with Illumina and to use a whole chromosome shotgun approach to sequence the 3DS/L chromosome to help assemble the physical map. Transcriptome sequencing is planned as well to provide additional resources to the community for the genome sequence annotation.

Session six afforded time for roundtable discussions that focused on four questions:

1) Can we use new technologies to sequence the wheat and barley genomes?
2) Which level of quality do we want to achieve for which use?
3) Given technology advancements, what foundational resources must be developed in the next few years to be able to utilize the new technologies?
4) What pilot studies should we undertake in the next year?

Before these questions were answered, the discussions centered on the goals of the genome projects and the quality of genomic information needed to inform improved breeding projects. C. Feuillet and Robbie Waugh (Scottish Crop Research Institute) brought up the point that regardless of what is decided, the product should be useful for the breeders for they are the ultimate customers of the genomic resources. It was agreed that good examples of how to utilize the sequence should be built into the project from the very beginning to ensure rapid application to breeders. There was a general agreement that we would have only one opportunity for sequencing the two genomes and, thus, the consortia should strive for a gold standard reference sequence. A number of intermediate steps were mentioned during the roundtable discussions as well. First, it was pointed out that additional resources for the wheat genome sequencing project would be needed. A consensus was reached as well on the need for new BAC libraries (2^{nd} enzyme and random sheared) from the whole Chinese Spring wheat genome to fill gaps that will be found in the chromosome-specific physical maps. Additional transcriptomic resources (EST, FLcDNA) will be needed also and can be generated now at reduced cost using the next generation sequencing technologies. The sequences of the other cereal genomes (rice, maize, sorghum, Brachypodium) will be instrumental as well in annotating the barley and wheat genomes. Finally, there is an agreement that bioinformatics platforms need to be developed to enable the wheat and barley genome sequencing projects. A database resource such as Ensembl for plant genomes was discussed and is planned at EBI. This database could then be used as a

central DB for linking the community databases and integrating all data. The need to build up mechanisms to involve the users was re-emphasized.

Overall, there was a consensus that the wheat and barley genomes could be sequenced; however, additional pilots need to be conducted to determine exactly how to approach these genomes particularly. Representatives from the next generation sequencing companies were asked to give an estimate for the cost of sequencing chromosome 3B (1Gb) for which a BAC MTP is available, as the first challenge and pilot project. Marcus Droege (Roche) presented a proposal with a combination of WGS on sorted chromosome and BAC by BAC sequencing. He stated that for WGS and 12X coverage the cost would be 18,000 € per 100 Mb, 10X coverage of 3 Kb inserts with paired ends would be another 6,000 €, 20X coverage of 20 Kb inserts would be 3,000 €, and thus the total for chromosome 3B (1Gb) for sequencing alone would be 270,000 €, not including labor, assembly and annotation. S. Humphray also expressed that Illumina is interested in being involved in a pilot on 3B with 10Gb Illumina paired end runs on sorted chromosome fragments. C. Feuillet asked whether a mixed approach can be discussed and if the assembly software can be adapted to accommodate sequence data from multiple platforms. This remains an open question and there was a consensus that this should be explored as well. Roche and Illumina expressed interest in sequencing sorted barley chromosomes as well.

Members of research funding organizations from Europe discussed funding opportunities during session seven and K. Eversole presented information from NSF about funding opportunities for plant genomes in the U.S. European presenters included Annette Schneegans, (European Commission), Rainer Büschges, (BMBF, Germany), Daniel Richard-Molard (Ministry of Higher Education & Research, France), Hélène Lucas (INRA France), Francis Quetier (ANR, France), and Joanna Jenkinson (BBSRC, UK). The different national grant agencies presented highlights of their research funding strategies that may be relevant to wheat and barley. At the EU level, while it is clear that there is unlikely to be any further opportunity for support within the framework of the COOPERATION program in the FP7 period, since the Triticeae Genome project is essentially the flagship project for structural genomics projects of these species for the next few years. However, there may be opportunities to use the CAPACITIES program to build up infrastructures needed to support wheat and barley genome sequencing. Finally, D. Richard- Mollard announced that the French Research Ministry and the ANR (represented by F. Quetier) have decided to provide a large grant to fund the sequencing of the wheat 3B chromosome, following the construction of the physical map.

Building from the discussions of the seven sessions, C. Feuillet presented a diagram showing the outline strategy and important steps that were agreed upon during the wrap up session. They include:

- A consensus was reached on a **two phase sequencing strategy** with:

- Phase 1: obtaining a good quality sequence of the wheat and barley genomes that can be used as soon as possible to develop tools for breeding and that represents a platform for phase 2. Pilot sequencing projects on chromosomes 3B and 3H of wheat

and barley will serve to establish the most cost effective approaches for the wheat and barley genomes. Roche 454 Titanium and Illumina Solexa technologies will be tested separately and in combination on sorted chromosomes and on the minimal tiling paths. Furthermore, the potential utility of WGS paired-end datasets on the diverse next generation sequencing platforms in pilot projects should be explored. Such datasets are needed to train algorithms for Triticeae genome characteristics and advance the approach. At the same time, such data will deliver "gene-catalog" sequence datasets that complement EST resources for marker and breeding-tool development.

- Phase 2: achieving high quality "gold standard" sequences that will enable all functional and structural analyses of the two genomes.

- To accomplish these goals, it was recognized that **new capacities and capabilities** need to be established. These include:

  - New whole genome BAC libraries from wheat (Chinese Spring) to fill gaps from the individual HindIII sorted chromosome BAC libraries. JIC and CNRGV proposed to construct new BAC libraries with another restriction enzyme (JIC) and from sheared DNA (CNRGV) for this purpose.

  - **New resources for annotation**: using new sequencing technologies, FLcDNA sequences, and deep transcriptome sequencing should be performed.

  - **Bioinformatics capacities** for handling the genome sequences: A **central database** (such as Ensembl, phytozome…) could be established to link the community databases and have a single point of entry for users and curators. E. Birney (EBI) is developing a strategy to extend the Ensemble platform to plants. A Triticeae annotation working group is already established (see IWGSC website tools page) and will lead these discussions.

  - **A platform to get feedback from the sequence users.** The importance of having the end users involved early in the process to ensure the quality and utility of the sequences is central to the success of the two projects and the achievement of the goals of the two consortia.

  - Sufficient **human resources** available and trained for the projects

- Develop **long-term, sustainable mechanisms for funding the sequencing and exploitation** of the wheat and barley genome sequences.

The meeting was adjourned.

IWGSC-IBSC Workshop on Sequencing Technologies
11-12 September 2008
Genoscope National Sequencing Center
Evry, France

Workshop Objective: Develop strategic roadmaps for sequencing the wheat and barley genomes

Agenda

Thursday, 11 September 2008

07:30 am – Shuttle departs from Novotel and All Seasons for transport to Genoscope

08:00 am – 12:00 noon – **Meeting Registration** *(Location: outside F. Jacob Room)*

09:00 am – 09:15 am -- **Welcome and introduction** (*Location: Conference Room*)

- Kellye Eversole, IWGSC
- Nils Stein, IBSC
- Patrick Wincker, Genoscope

09:15 am – 09:20 am – **Overview of workshop goals**

- Jane Rogers, BBSRC

09:20 am – 09:45 am – **Session 1. Genome structure of wheat and barley**
Objective: Provide summary of work in the last decade, requirements of the genome sequence, and challenges.

- Catherine Feuillet, INRA

09:45 am – 10:30 am. **Session 2. Experience from other genome projects**
Objective: Discussion of strategies, challenges, results, utility; lessons learned that should be considered for de novo sequencing of wheat and barley. Moderator: Jane Rogers.

- ***Panel on complex, non-plant agricultural species projects***
  - Jane Rogers, BBSRC, Human and Porcine Genomes
  - André Eggen, INRA, Bovine Genome
- Discussion

10:30 am – 10:45 am. **Coffee break** (*Location: F. Jacob Room*)

10:45 am – 12:45 pm. **Continuation of Session 2.** Moderator: Jane Rogers.

- ***Panel on economically important sequenced plants***
  - Francis Quetier, ANR, Rice Genome Sequence
  - Dan Rohksar, DOE-JGI, Soybean & Sorghum Genome Sequences
  - Patrick Winker, Genoscope,Grape Genome Sequence
  - Pat Schnable, Iowa State University, USA. Maize Genome Project
- Discussion

12 :45 pm – 02 :00 pm – **Lunch** (*Location : F. Jacob Room*)

02 :00 pm – 04 :00 pm – **Session 3. Current wheat and barley projects**

Objective: Results of ongoing wheat and barley de novo sequencing pilot studies using new technologies alone or in combination with Sanger. Moderator: Catherine Feuillet.

- Nils Stein, IPK, Pilot projects on barley
- Jean-Marc Aury, Genoscope, Pilot project on sequencing wheat BACs
- Etienne Paux, INRA, Wheat pilot projects
- John Fellers, USDA-ARS, KSU, Methyl filtration projects
- Discussion

04:00 pm – 04:30 pm – **Coffee break** (*Location : F. Jacob Room*)

04:30 pm – 06:30 pm – **Session 4. New Sequencing Technologies**

Objectives: Overview of the application of sequencing technology platforms to sequencing large, complex genomes and, in particular, how the technology can be applied to wheat or barley. Specific topics include: Physical mapping, sequence generation, sequence assembly, QC / QA. Moderator: Jane Rogers.

- Lei Du, 454 Life Sciences Roche - 454 GS FLX
- Sean Humphray, Illumina -Genome Analyzer
- Avak Kahvejian, Helicos Biosciences – Single Molecule Sequencing
- Marco Pirotta, Applied Biosystems – SOLiD System
- Questions and discussions.

06:30 pm – **Shuttle transportation to Novotel for workshop reception and dinner**

Shuttle service will be provided from Genoscope to Novotel Hotel only

07:15 pm – **Reception at Novotel** (*3 Rue de la Mare Neuve, Lac de Courcouronnes)*

08:00 pm – **Dinner at Novotel**

Following the dinner, a shuttle will provide transportation to All Seasons for guests staying there.

Friday, 12 September 2008

07:30 am – Shuttle departs from Novotel and All Seasons for transport to Genoscope

08:00 am – 10:30 am – **Meeting Registration** *(Location: outside F. Jacob Room)*

09:00 am – 10:15 am – Session 5. **Bioinformatics** *(Location: Conference Room)*
Objective: Discussion of bioinformatics needs and opportunities related to sequence assembly, map and sequence integration, and database submission (i.e., how to accommodate the data and in which form). Moderator: Nils Stein.

- Richa Agarwala, NCBI, NIH, USA
- Dave Edwards, ACPFG, Australia
- Discussion

10:15 am – 10:30 am – **Coffee break** (*Location: F. Jacob Room*)

10:30 am – 12:30 pm – Session 6. **Roundtable discussions**
Objective: Discussion of key questions for moving forward. Moderators: Jane Rogers, Catherine Feuillet, and Nils Stein.

- Can we use new technologies to sequence the genomes?
- Which level of quality do we want to achieve for which use?
- Given technology advancements, what foundational resources must be developed over the next few years to be able to utilize the new technologies?
- What pilot studies should we undertake in the next year?

12:30 pm – 02:00 pm – **Lunch** (*Location: F. Jacob Room*)

02:00 pm – 03:15 pm – Session 7. **Research funding**
Objective: Discussion of potential opportunities for funding by workshop sponsors and government funding agencies. Moderator: Kellye Eversole

- Annette Schneegans, European Commission
- Rainer Büschges, BMBF, Germany
- Daniel Richard-Molard, Ministry of Higher Education & Research, France
- Hélène Lucas, INRA, France
- Francis Quetier, ANR, France
- Joanna Jenkinson, BBSRC, UK

03:15 pm – 04:00 pm – Session 8. **Wrap-Up and next steps**
Objective: Discussion of next steps and whether a follow-up workshop or broader conference should be planned for 2009. Moderators: Jane Rogers, Catherine Feuillet, Nils Stein, and Kellye Eversole.

04:00 pm. **Adjournment**