Part 1A Track Record

This proposal brings together scientists with relevant genomics and bioinformatics expertise to make a major UK contribution to the international effort to generate a reference sequence of the bread wheat genome, to access gene variation in multiple wild relatives, ancestral species and mutant lines, and to make sequence data, analyses and toolboxes for analyses available to users world-wide.

Principal Investigator

Dr Jane Rogers is the Director of The Genome Analysis Centre (TGAC). She will act as coordinator of the LoLa programme in wheat genomics. TGAC is a new institute of the BBSRC, established in 2009 in Norwich through a partnership with the East of England Development Agency and Norfolk Local Authorities to further genomics research and application through high throughput sequencing and bioinformatics, primarily in plants, animals and microbes. TGAC operates 2nd generation sequencing instruments (currently (Roche 454 Titanium (x2), Illumina GA2 (x2) and a Hi-Seq) and Applied Biosystems SOLiD 4 (x2)) with associated computing infrastructure (0.5Pb mirrored storage and a high performance cluster (4 x 256Gb RAM) for genome sequence assembly, comparative analyses and annotation. In its first year of operation, TGAC has applied next generation sequencing to a variety of genome sequencing projects ranging in size from bacterial genomes (typically 4-7Mb) to plant genomes of 500Mb (red clover) and 2Gb (rubber tree). TGAC is working with JIC and EBI on the sequence assembly of wheat chromosome arm 3DL. It was also recently commissioned by the International Wheat Genome Sequencing Consortium (IWGSC), to undertake Illumina survey sequencing of flow-sorted wheat chromosome arms for approximately 40% of the genome and to coordinate the data quality assessment and assembly efforts with contributions from other international partners and the IWGSC Bioinformatics Board led by Klaus Meyer (MIPS).
Prior to establishing TGAC, Dr Rogers was a founder member of the Wellcome Trust Sanger Institute and sat on the Board of Management for 14 years (1993-2007), with responsibility for the high throughput sequencing facility. Over that period the Institute made major contributions to large vertebrate genome projects: human (**Nature 409, 860**; **Nature 431, 931; Nature 437, 1299**), mouse (**Nature 420, 520**), zebrafish, pig (BBA16309) and gorilla, that are all represented in Ensembl and support genomics and downstream applications projects worldwide; they provided UK sequence contributions to *Arabidopsis thaliana*, *Medicago truncatula* (BBSB11648), tomato (BBC5193701) and *Hyaloperonospora arabidopsis* (BBC5212441) genome sequences; and through the Pathogen Sequencing Unit delivered genome sequences for over 100 small genomes (bacteria and eukaryote pathogens). Many of the projects involved large international consortia, requiring management and coordination at strategic and delivery levels to achieve the production of resources with high utility and usability to their respective research communities. Dr Rogers is a member of the IWGSC Coordinating Committee. Time commitment 20%.

Co- Principal Investigators

Dr Mario Caccamo is the Head of Bioinformatics at TGAC. He will lead the development and application of assembly and genome analysis bioinformatics.
His research interests focus on the development of efficient algorithms and software tools for the assembly and annotation of genomic sequences. Dr Caccamo previously worked at the Wellcome Trust Sanger Institute in the genome projects for both the model organism *Danio rerio* (Zebrafish) and *Sus scrofa* (Pig). In these roles he gained considerable experience in aspects of sequencing, assembly and annotation. In particular, he designed and developed software tools to assist the integration of the genomic sequences from different sources (i.e. BAC sequencing and whole genome shotgun assemblies) to improve the quality of genome references. Dr Caccamo joined the European Bioinformatics Institute in July 2007 to work in the development of the European Genome-phenome Archive (EGA). The aim of this project was to implement a public repository for clinical data that is subject to consent agreement. He also represented the EBI in the Genome Reference Consortium (GRC), this group’s activities are centered around the implementation of the tools and data standards required for the maintenance of the high-quality genome references

including the human and mouse sequences. The wheat genomics effort is integral to Dr Caccamo's research interest in developing methods for *de novo* assemblies and he is co-author of Cortex, a new memory-efficient algorithm (ms submitted) that is being used for large eukaryotes genomes. He also takes a more general interest in developing tools that enable biologists to take advantage of the data generated by the latest sequencing technologies. As Head of the Bioinformatics division at TGAC Dr Caccamo is experienced in supervision and management of staff working in the administration and execution of the informatics pipelines at TGAC. Dr Caccamo is a member of the IWGSC Coordinating Committee and serves on the IWGSC Bioinformatics Board. Time commitment 10%.

Dr Michael Bevan is a Project and Programme Leader at JIC. He will lead the chromosome physical mapping and whole genome reconstruction activities and contribute to genome re-sequencing, sequence annotation and analysis.
The Bevan group's current expertise lies in sequence analysis, database provision, genetic and physical mapping and project coordination in plant genomics research. The Bevan group has made several important contributions to plant genomics: they sequenced the first plant chromosome (**Nature 402, 769**), initiated and led the completion of the Arabidopsis genome in 2000 (**Nature 408,796**). They developed the largest transposon tagging population for Arabidopsis of 27,000 lines as a key resource for functional genomics. They developed one of the major functional genomics databases and web sites representing all functional genomics data and gene expression patterns that is now widely used, including by industry. They made physical maps of rice chromosome 2 as part of the International Rice Genome Sequencing Consortium, and were authors on the first paper sequencing a grass genome (**Nature 436, 793**). Dr Bevan was a co-PI on a DOE Joint Genome Institute genome project sequencing the genome of the grass *Brachypodium distachyon*, recently completed and published (**Nature 463, 763**). This genome sequence, the first of a pooid grass, provided the foundation for assessing gene synteny in grasses. This approach to ordering genes is proving to be very productive in barley and wheat. Bevan is currently a Partner in the EC-funded Triticeaegenome project, in which his group is making a BAC-based physical map of chromosome 3DL. He is also sequencing chomosome 3DL as part of a BBSRC T&R project with EBI and TGAC aimed at establishing next generation sequencing and assembly methods for applying to all wheat chromosomes. He is also a PI (with Edwards and Hall) working in a BBSRC-funded project aimed at using next generation sequencing to discover sequence polymorphisms in wheat. This project has produced the most extensive sequence data set for bread wheat and 4 elite breeding lines, and lays the foundations for more extensive genome sequence. This wheat genomics work is directly relevant to JIC's world-leading work in wheat genetics, trait analysis, gene identification and germplasm improvement. It has also established key links with TGAC and EBI to form the foundation for systematic sequencing of the wheat genome. Dr Bevan is a member of the IWGSC Coordinating Committee. Time commitment 5%.

Dr. Cristobal Uauy is a Project Leader at the John Innes Centre (JIC) who will lead the re-sequencing activities, wheat functional genomics and provide wheat genetic expertise to the project. Dr Uauy led the map-based cloning of the first QTL in polyploid wheat, *GPC-B1*, a transcription factor regulating senescence and with pleiotropic effects on grain mineral concentration (**Science 314,1298**). He also identified *Yr36*, a resistance gene which confers partial and broad spectrum resistance to wheat yellow rust (**TAG 112, 97**) and completed the positional cloning of *Yr36* (**Science 323,1357**). To validate *Yr36*, he developed wheat EMS-TILLING populations and alternative detection methods (**BMC Plant Biology 9,115**; *Highly accessed*) to rapidly access variation across candidate genes. These TILLING populations were also screened to identify knockout mutations in the *GPC* homoeologues to extend grain filling duration (by delaying senescence) and to produce wheat lines with lower protein content, an important breeding target for soft varieties used in biscuit making. These mutants are currently being used by KWS and Limagrain in their programmes, exemplifying how TILLING offers the possibility to develop useful genetic variation for breeders. Together with Andy Phillips at RRes, he was recently awarded a one year BBSRC-BBR grant (BB/I000712) to develop publicly available TILLING populations for the UK plant science community and to investigate novel approaches for TILLING, including genome-capture technology followed by sequencing. The high mutation density made possible by the polyploid nature in wheat means that these populations have a combined mutation frequency of ~65 mutations per kb, almost two orders of magnitude higher than the levels of

natural variation in cultivated wheat. This translates into a probability of >90% of identifying knock-out alleles for any given gene in wheat, effectively changing the paradigm of what can be done in functional gene analysis in wheat. These populations will form the basis of Objective 2.2.
Dr. Uauy is a co-PI on the wheat Pre-Breeding LoLa application, where he is developing chromosome-segment substitution populations of wild emmer and *Aegilops tauschii*. He is also working closely with the major UK wheat breeders in a large BBSRC-defra-HGCA funded LINK project (BB/I01800X) to investigate the genetics of preharvest sprouting in UK-adapted material. His involvement in these projects offers a good bridge between the researchers in this proposal, the UK breeders, and the wider international wheat research community. Time commitment: 5%.

Andreas Magusin is a Senior Scientist at JIC responsible for statistical genomics applications. He will be responsible for developing bioinformatics applications for wheat molecular breeding, gene identification and QTL analysis. He has developed applications using the Galaxy package that is used in Arabidopsis for chromatin IP analysis in Arabidopsis (Nature 462, 799) and Brassica genome sequence variation (**Molecular Breeding 26, 91**; **BMC Plant Biology 9, 50**).

Dr. Andy Phillips leads a multidisciplinary group at Rothamsted research working on hormone signalling in plants and its applications to wheat. This work is supported by BBSRC grants from Agri-Food, the Crop Science initiative and the LINK programme. Within a CSI project (BB/E0069221), in collaboration with a team at JIC they are investigating alternative dwarfing genes in wheat through mapping GA biosynthetic and signalling genes and correlating these with QTLs for plant stature, as well as investigating the interaction between the Rht dwarfing genes and drought stress. In a LINK project (BB/D0073481) that also involves several academic partners (including Prof. Holdsworth at U. Nottingham) a well as most of the UK wheat breeding companies, they are investigating the genetic and physiological basis for variation in Hagberg Falling Number (HFN), a measure of wheat quality that relates to α-amylase levels in the grain. Within this project they have large scale field experiments on multiple sites which is enabling them to identify genetic loci that protect against low HFN. A major causes of loss of grain quality in wheat is pre-harvest sprouting, which has led them into investigating the role of coat-imposed dormancy in a current application. In project BB/D019001 they are determining the role of gibberellin in controlling grain size in wheat, where they have produced transgenic lines that over-produce bioactive GA in the embryo and have 15-20% larger grain (Phillips 2006 Patent number WO2006032916). In addition, within this project they have developed antisense lines designed to suppress GA biosynthetic and signalling genes in developing grain, which has generated lines with a range of grain phenotypes. Dr. Phillips was also a programme leader within the Defra Wheat Genetic Improvement Network (WGIN), in which he collaborated with Dr. Kim Hammond-Kosack (RRes), and Prof. John Snape (JIC). Within this programme, TILLING was established as a technology for the identification of natural and induced variation in hexaploid, tetraploid and diploid wheat. EMS-mutagenized TILLING populations of wheat have been generated and a number of targets involved in wheat growth and development and pathogen resistance have been screened (**J Exp Bot 60 2817**; **Mol Breeding 25,145**). The TILLING platform has been semi-automated on a liquid handling robot bought on a BBSRC REI grant and contracts for services have been completed from academic collaborators. In addition, a grant from the Tools and Resources Development Fund (BBE0251611) has enabled the establishment of a novel technique for mutation discovery, High Resolution Melt Analysis. This has proved to be very sensitive in identifying point mutations within the EMS-mutagenised TILLING population of wheat developed in the WGIN programme and has been integrated into the workflow for mutation discovery. This technique has been used to identify null mutations in all three homoeologues of the *SbeIIa* gene from wheat (Botticella et al., unpublished).

Dr Paul Kersey leads the Ensembl Genomes team at the European Bioinformatics Institute (EBI)**.** The EBI is an academic research and part of the European Molecular Biology Laboratory. The EBI provides freely available data and bioinformatics services to all facets of the scientific community in ways that promote scientific progress; contributes to the advancement of biology through basic investigator-driven research in bioinformatics; provides advanced bioinformatics training to scientists at all levels, from PhD students to independent investigators; and help disseminates cutting-edge technologies to industry. The Ensembl Genomes team develops services for non-

vertebrate genomes as part of the nucleotide section of Protein and Nucleotide Database (PANDA) Group, which provides the EBI's nucleotide- and protein- based services (including the European Nucleotide Archive, UniProt, InterPro, and Ensembl). Ensembl is a sophisticated infrastructure for genome analysis and visualisation originally developed in the context of the human genome project. Since April 2009, it has been used by Dr.Kersey's team to provide access to genome scale data for bacteria, fungi, protists, plants and invertebrate metazoa, with 7 public releases so far. The team has also been heavily involved in developing resources for genomic variation (having developed variation resources for *Arabidopsis thaliana*, rice and grape), and is developing software for the *de novo* assembly of large genomes from next-generation sequencing data.
The team has recently been funded to work on the annotation of plant pathogens by the BBSRC; and from other sources, on the wheat stem rust pathogen, the model yeast *Schizosaccharomyces pombe*, the annotation of bacterial pathways, invertebrate vectors of human diseases, and the scuttle fly. The development of Ensembl Plants is an international collaboration with Gramene in the U.S.A.; the EBI has received BBSRC funding to support the development of resources for genomic variation in *Arabidopsis thaliana*, on developing algorithms for genome assembly; and for international travel to support the development of plant genome databases

Contribution to UK economic competitiveness
The potential impact of the project will be maximized by the strong commitment, focus and expertise of the research team. We have substantial records of achievement in genomics, genome assembly, database informatics and wheat research as described above. Dr Rogers also has experience of leading these programmes: she was a major player in the Human Genome Project, that established the Wellcome Trust Sanger Institute and the UK as a world leader in human genomics and its application to the understanding and treatment of disease. The proposed research will be carried out in state of the art genomics labs and directly embedded in the wheat programmes of two BBSRC institutes to ensure relevance and impact.

Part 1B. Data sharing

The aim of this proposal is to contribute to the international effort to generate a reference genome sequence of bread wheat variety Chinese Spring 42, to re-sequence the gene space of mutagenised populations, and to re-sequence the gene space of multiple Triticeae genomes relevant to wheat pre-breeding research. Sequence reads from the reference genome and from re-sequencing will be deposited in the European Nucleotide Archive as soon as possible in accordance with the principles agreed at the Toronto International Data Release Workshop (2009)(1), and the analysed reference genome sequence will be made publically available via an Ensebl database at EBI. Access to analyses of sequence variation will be available through EBI and local databases as soon as practicable after completion of analyses.

Data areas and data types*:* The main data type generated will be DNA sequence reads and assemblies. Raw sequence data, generated by next generation sequencing machines, will be stored temporarily at TGAC for quality assessment before submitting them to the short read archive at EBI The sequence files (FASTQ format) will be aligned, assembled and annotated to identify genes and other features.

Standards and metadata*:* Sequence data and assemblies will be stored in resilient backed-up file servers at TGAC and EBI. Analysed data will include predicted sequence polymorphisms (including mutations), the predicted proteomes of wheat and barley, wheat gene family relationships and comparative genomics data with the genome zipper of *Brachypodium*, sorghum and rice. Standard formats (e.g. FASTA, ENA flat file, and in Ensembl data schemas) will be used wherever available.

Relationship to other data available in public repositories*:* This project will substantially increase the amount of genome sequence for wheat and complement other international efforts focused on chromosome-based efforts to produce a high quality reference sequence with high utility. It will provide a significant resource of sequence variation in agriculturally- relevant Triticeae, and provide the means to identify polymorphisms for genetic marker development and gene function analysis. As such the data will be an important new community resource and will be a major contribution to the foundation for genomics-led wheat breeding and gene discovery.

Secondary use*:* The sequence data and assemblies will be used by the research community and the wheat breeding industry to develop new genotyping platforms, to implement at a larger scale molecular breeding methods to improve the wheat crop worldwide, and to expand knowledge of wheat gene function.

Methods for data sharing*:* Sequence data and assemblies will be submitted to EBI depositories as soon as practicable after generation. The emerging wheat genome sequence will also be included in the Ensembl Plants database at EBI, with access provided through a variety of programmatic and interactive interfaces (i.e. web browser, FTP, Perl API, public database server (MySQL), query-optimised data warehouse, DAS servers). Sequence variation data will be held at EBI and analysed via local database resources and toolboxes held at TGAC or the JIC data centre. These interfaces provide the diverse user community with practical ways of using the sequence data generated. Publications of data analyses will be made throughout the project and after it. Seeds of mutant lines will be available through requests to the JIC Germplasm Resources Unit. Our previous work exemplifies our commitment to making our resources and information publicly available.

Proprietary data*:* All sequence data generated in this project will be deposited in public databases as soon as practicable after generation and assembly. None of the generated data, nor the results of automated analyses, such as predicted polymorphisms, will be proprietary to the researchers.

Timeframes*:* Sequence data and assemblies will be released during the course of the project. Genome- wide sequence variation and gene mutations will be released as soon as practicable after analysis. Sequence assembly algorithms and applications will be published during and after the project.

Format of the final dataset*:* The methods and global analysis of the data generated in this project will be published in journals. Analysed, genome-wide data will be accessible through an Ensembl browser, which will also enable customization by individual users and data downloads to other databases.

Part 2

Summary

Having a high quality reference sequence of the bread wheat genome will significantly accelerate progress in breeding and gene discovery in this important global food crop and hence will make a key contribution to achieving food security. This project aims to integrate and coordinate UK expertise in genomics and bioinformatics to contribute to sequencing the genome of hexaploid bread wheat as partners in the International Wheat Genome Sequencing Consortium (IWGSC). The project will also access genetic variation in wheat and the genomes of other Triticeae to enable more efficient genomics- led breeding and biotechnological applications. The information generated will be provided in useful databases and interface formats for breeders and scientists worldwide as soon as practicable after generation and analysis.

The specific objectives are to:

- define a set of high quality, non-redundant sequences representing >90% of genes for bread wheat Chinese Spring 42;
- produce physical maps and sequence four chromosomes to high quality reference standards;
- annotate and analyse predicted genes;
- establish long-range scaffold methods for ordering genes and apply this at chromosome and whole genome levels;
- re- sequence genes of mutagenised populations;
- re-sequence genes of diverse Triticeae genomes relevant to wheat pre-breeding research;
- create databases and bioinformatics tools for applying the sequence resources to crop improvement;
- maximize the impact of the research through training and outreach.

Strategic relevance of the project to BBSRC

Securing food supply on a global scale requires rapid solutions to a complex set of unprecedented challenges, including rising demand due to rapid population increases and social mobility, global climate change, rising energy costs, and land, water and nutrient limitations. Finding and implementing these solutions is a top priority for governments and scientists worldwide, and has been articulated as a key strategic objective by BBSRC. Opportunities for plant science to contribute to global food security include increasing the yield and quality of crops, combating diseases, achieving yield stability in sub-optimal growing conditions, and increasing maximal yield potential. Utilizing non- food components of food crops, such as cell wall material and by- products of food production to produce energy and industrial feed-stocks, also has a role in reaching sustainability and maximizing overall yield of renewable resources from limited land and soils.

Grass crops are essential for human existence by directly and indirectly serving as the primary source of human nutrition. Wheat, rice and coarse grains such as maize are the most important human food crops; therefore securing future grain supplies from a sustainable production system is a critically important strategic and scientific objective. Wheat is the main arable crop in the UK, planted on 43% of arable land (average 2005-2009), with an annual farm gate value of ~£2.5bn and a processed product value of approximately £150bn. In 2009/10 global wheat production is estimated to be 682 m tonnes, and it is the most widely traded agricultural commodity with a 2007 value of $35bn (2). Current trends indicate increased production from China and increased global demand against a background of long- term production problems in major growing areas caused by drought, among other constraints. Looking forward, new threats to production such as rust epidemics add further uncertainty to the prospects for food security.

Yield increases in wheat are slowing compared to past gains achieved primarily through genetic and agronomic improvement and also in relation to other grain crops, notably maize (3). To meet predicted large increases in demand and greater yield stability, new breeding programmes aimed at incorporating a far wider range of genetic diversity into wheat breeding are underway, including large coordinated projects for wheat genetic improvement at CIMMYT (International Maize and

Wheat Improvement Center), in France, the US and in the UK. These form part of the Global Food Security Programme, that is led in the UK by the BBSRC. In parallel, intensive research in many aspects of plant biology, much of it carried out in experimental species such as Arabidopsis, has generated deep knowledge of genes involved in traits such as environmental adaptation, nutrient acquisition and downstream modulators of disease response, all of which are directly relevant to crop plant improvement. Nevertheless there is currently a gap to be bridged between plant research and wheat improvement.

Genomics is a major and significant component of modern biology and biology- based industries. A high quality reference sequence of wheat will provide access to the relatively compete gene catalogue, the regulatory elements that control function and a framework for understanding genomic variation (4). Therefore access to a high quality wheat genome sequence will lower barriers to crop improvement and release the full potential of plant science to meet the challenges of food security and sustainable agriculture. For example, wheat genome sequences, when delivered to users through databases, browsers and specialized bioinformatics tools, can be used to identify genetic diversity underlying useful traits, to reveal mechanisms driving genome change, and to integrate knowledge of gene function across diverse organisms. In this way genomics and associated bioinformatics analyses provide a key framework for distributing and sharing knowledge between academia and industry. In crop plants with genome sequences (e.g. rice and maize) (5), genome wide association studies (6) are revealing extensive genetic variation underlying key traits. This research provides a systematic way to uncover the genetic basis of agronomic traits and a wealth of informative markers for genotype selection in marker assisted breeding and candidate gene identification. In this way genomics and genome bioinformatics is dramatically accelerating the improvement of these crops. The absence of a reference genome sequence of wheat is currently limiting similar routes for improvement and increased sustainability of this globally important crop.

Over the last 5 years, the IWGSC has established a strategic roadmap to achieve a high quality reference sequence of the wheat genome.This proposal describes a UK component of the international project that will make decisive and innovative contributions to producing a reference sequence for the bread wheat genome using cost-effective approaches and within a timeframe supporting germplasm improvement and trait biology projects. The outcomes of this project include informative gene-based polymorphisms in a complete genome framework to accelerate wheat breeding programmes, whole genome resources that can identify an entirely new and essentially complete spectrum of genetic variation for analysis and incorporation into breeding and transgenic improvement programmes, and an informatics infrastructure supporting both industrial application of the wheat genome sequence and the translation of basic research into wheat improvement.

Strategic relevance of the project to the goals of the host institutions

The three partners have a strong commitment to the establishment of a strong wheat genomics programme in the UK and its applications for crop improvement because they are centrally important to our long- term research objectives. Our commitment is demonstrated by a productive track record of working together in the field of plant genomics and contributing jointly and separately to major international research programmes. The wheat genome project is a major platform for the application of TGAC's research in the development of novel sequencing technology, including new sequence assembly methods, gene identification and annotation, and other bioinformatics tools for plant research. Tackling and representing the complexity of the wheat genome will require research and development in all these components and will provide a strong foundation for TGAC to continue to contribute to and lead key strategic programmes in food security and sustainable production. With a strong contribution to the genome sequence, TGAC will be well positioned to help in the coordination of the sequencing and analysis efforts of international partners in the IWGSC and to ensure the delivery of a product with uniform quality across the genome that constitutes a valuable long-term resource.

Wheat research is centrally important to the long- term research objectives of JIC, TGAC and RRes in the fields of food security and sustainable production. For example, access to a high quality reference wheat genome sequence and sequence variation in Triticeae will revolutionize research in pre-breeding, facilitate the identification of genes underlying traits, and facilitate translation of research from models to crops. The wheat genome will also promote novel bioinformatics research at JIC, RRes and TGAC to exploit the genome sequence and to facilitate uptake by industry, in addition to helping train and recruit the next-generation of crop scientists. The dependence of future goals in plant research on progress in wheat genomics and its applications are laid out in the Institutes' ISPG applications for 2012-2017. At JIC the GRO and BIO Strategic Programmes have approximately 50% of researcher time aligned to wheat research, and at RRES wheat is the main focus of activities in the 2020 Wheat and Designing Seeds Programmes.

The provision of open access to molecular biology data, in as useful a form as possible, is the core mission of the EBI. Genomic sequence is not only interesting in its own right, but provides an organizing principle for other "-omics" data. Following the successful development of the (vertebrate-focused) Ensembl project, the EBI has committed resources to the application of Ensembl technology to a broader range of species, to provide a uniform interface for access to genome-scale data in accordance with the needs of scientific communities. Currently 8 staff are employed from EBI core funds in technology development, release production and service maintenance (for all non-vertebrate species). These are supplemented by additional staff employed to work on specific community-driven projects in particular areas. The development of Ensembl Plants is of particular importance given the increasing awareness of the imminent shortage of food and fuel, and we have developed (or are developing) strategic collaborations with Gramene (in the United States) and leading groups from the European Union in a bid to construct an international resource for plant genomic data.

Timeliness

The size and complexity of the wheat genome have been major barriers to progress in sequencing this important crop genome and the large genomes of other major grass crops, for example barley, oats and energy grasses. Two major technological advances have come together in the past 2-3 years, however, that now make it feasible and cost-effective to take on the challenges posed by a large, hexaploid genome: 1) The emergence of next generation sequencing technologies with capacity to deliver hundreds of Gigabases of sequence in a single run for an overall cost that is reaching less than $10,000 per human genome has driven the development of novel computational methods for sequence assembly and analysis of short read sequence data; and 2) the development of a strategy coordinated by the IWGSC based on the physical mapping and sequencing of individual chromosome arms flow-sorted from aneuploid lines of Chinese Spring 42 provides a method for assigning annotated sequences and particularly homeologous genes to individual chromosomes. The initiative to map and sequence chromosome 3B led by Catherine Feuillet at INRA (5), has demonstrated the feasibility of the approach, which is now being used for all 21 bread wheat chromosomes through the IWGSC. Recent progress in barley genome analysis (N. Stein *pers comm.*) also provides a very clear example of what could be achieved in wheat. Within the next twelve months, building on initiatives funded by BBSRC to generate genome-wide 454 sequence of Chinese Spring and Illumina survey sequences of individual chromosome arms undertaken by IWGSC members, the first assignment of gene sequences to individual chromosomes will be made. It is crucial for the utility of these resources and their application to crop improvement programmes that these initiatives are extended to the generation of a relatively complete gene set with assignment of genes to their chromosomal locations across the genome in as accurate a sequence context as can be achieved, to accelerate efforts to understand their regulation. It is thus extremely timely for the UK to contribute to and provide leadership in this international effort and to establish resources for applying genome-wide data to new wheat breeding initiatives. Whilst the genome is ready to be sequenced, nevertheless, several technological innovations and adaptations will need to be established to achieve a high quality reference genome sequence, and these are described here. This project will contribute to acceleration of progress in wheat crop improvement, a programme that has become a strategically important objective of funding agencies worldwide, as, governments are faced with the challenges of feeding a growing world population.

Training

The reference genome sequence data and Triticeae sequence variation data will form one of the largest and most complex sequence datasets generated to date. We aim to ensure these voluminous and complex data are provided in user-friendly formats to users, in particular industrial users who have not used genomic data routinely- for example in breeding. To ascertain user needs, specific workshops will be held to introduce the data and its potential applications. From this, tailored training programmes for different users will be developed to meet the identified needs. In particular we aim to impart the knowledge and skills required to apply wheat genomics and bioinformatics for crop improvement, with an emphasis on developing resources, tools and tested approaches that will facilitate the adoption of new methods. The project offers outstanding opportunities for training in genome bioinformatics to the post-doctoral researchers working on the project directly, and throughout the world. These include research in novel sequence assembly algorithms that incorporate third generation sequencing and long-range scaffold sequence, integration with physical maps, comparative genomics, and gene identification, annotation and gene family relationships. Developing databases and analytical tools that integrate field phenotype data with genome wide association as predictive breeding tools provides further distinctive training opportunities. Training in wet-lab research involving the application of genome selection for re-sequencing will also be provided.

Research Programme

Background

Bread wheat has an exceptionally complex genome comprised of three independently- maintained genomes each of which is approximately 6 Gb- more that the entire human genome (7). The three independent genomes of bread wheat are the A genome, derived from *Triticum urartu*, the B genome and the D genome from *Aegilops. tauschii* (8). The AABB genotype was formed by an initial hybridization approximately 0.5 mya, followed by another hybridization with *Ae. tauschii* approximately 10,000 years ago. The descendents of this initial hybridization event were then selected and bred to form modern bread wheat. Recent sequence analysis of a small region of wheat chromosome 3B (5) shows that genes are found predominantly as small (1-4) clusters, with an average density of between 1 gene/86 kb in proximal regions and 1gene/180 kb in distal regions of the chromosome. Genes and gene islands are separated by extensive tracts of nested retrotransposon repeats comprising approximately 85% of the genome. The gene content of diploid grasses is approximately 30-35,000 (9) suggesting bread wheat has approximately three times this number of genes. However, larger genomes are predicted to contain a proportionately larger number of genes and pseudogenes (7,10), leading to upper estimates of approximately 150,000 genes. These widely different assessments of gene number indicate that gene annotation in wheat requires careful attention to ensure that pseudogenes are classified correctly.

There are two features of wheat chromosomes that facilitate a systematic approach to this apparently daunting genome. First, aneuploid lines have been exploited to purify entire chromosome arms for BAC library preparation and direct sequencing (11). This reduces the scale of genome sequencing and physical mapping to an average of 500 Mb and permits the direct allocation of sequenced genes to specific chromosome arms. A landmark study utilizing chromosome purification (12) described the first BAC-based physical map of wheat chromosome 3B. This physical map, which now covers nearly 1000 Mb with 1283 contigs anchored by approximately 4000 markers, is the foundation for a BAC- based approach to chromosome sequencing; currently 12% of the chromosome has been sequenced.

The second feature facilitating sequence analysis is the high degree of conserved gene order in grass genomes. This reflects ancestral gene order that can be used to establish the approximate order of a significant proportion of wheat genes by alignment with the sequenced genomes of diverse grass sub-families, specifically rice and sorghum (5), and the Triticeae sub-family member *Brachypodium distachyon* (9). This approach, dubbed the “genome zipper” (13), when integrated with other information such as genetic maps and chromosome arm location has recently been shown to order approximately 75% of the estimated 35,000 genes in the barley genome (14). Such a synteny-based framework can provide an exceptional density of landmarks for gene identification and molecular breeding and is immediately useful. Locating non-syntenic genes, which are predicted to be rapidly evolving, and defining the precise order and sequence distance between genes, requires a long-range sequence-based scaffold for localization. This can be provided by a physical map of BACs, either by directly sequencing and assembling BACs, or integrating whole- genome and whole-chromosome sequencing with BAC sequencing. Alternative approaches to generating a long- range sequence scaffold (50- 100kb) could be taken using new sequencing technologies (15), but this requires considerable technical development and new assembly procedures.

A wealth of untapped genetic variation exists in wheat's ancestral species (16); including resistances to pathogens, environmental stresses such as drought (17), grain nutritional content, among others. The bread wheat gene pool is being extended and enriched by advanced breeding programmes. For example, the D genome diversity is being expanded by incorporating novel genetic diversity from *Ae. tauschii* through synthetic wheat production. This research is part of major global programmes coordinated by CIMMYT, to which the BBSRC makes major contributions. Therefore genome sequencing of ancestral species, and of hexaploid wheat landraces existing prior to modern breeding, will access genetic variation to aid marker-assisted breeding and gene characterization. Mutagenised populations (for example TILLING populations) are another key source of genetic variation needed for characterizing gene function in wheat as these have been shown to have mutation rates almost two orders of magnitude higher than the levels of natural variation in cultivated wheat (18). Sequencing the genes in these populations provides direct access to mutated genes for functional genomics studies.

Strategy

We aim to take a step-wise approach to achieving the final goal of a wheat genome assembly that builds on recent progress in both wheat and barley genomics. This approach ensures that data are provided to users in a timely way to meet the urgent needs of the research and industrial community for access to a majority of the wheat gene sequences and sequence variation. For example, one milestone is to provide the accurate sequence of a majority of the genes within a syntenic framework by the end of the first year of the project. This will provide users with genomic resources while subsequent, more time-consuming and technically complex work to establish a more accurate gene order in the high- quality draft genome sequence is conducted. In this way the utility and impact of the project can be maximised. This step-wise approach also enables the project to adapt to relevant technology developments, such that it is not entrained to specific technologies available at the start of the project. For example, spanning large genomic regions with sequence landmarks is required to order wheat genes accurately, but sequencing technology with this capability is still being developed. Similarly, assembly algorithms and the high capacity computational infrastructure needed for their operation are in continual development, so the project will be sufficiently adaptable to take advantage of new approaches while maintaining progress towards a clear set of final deliverables.

Summary of progress to date in current collaborative research

In the past year the partners and our collaborators have made important progress in developing cost-effective ways to sequence and assemble the wheat genome. This progress is described in the appended progress reports. In a BBSRC-funded grant (BB/G013985) 5x coverage of the CS 42 reference genome with long 454 reads has been completed and assembled. The non-repeat component of the sequences has been assembled into 5m contigs (n50 1028bp, largest contig 21kb) with 5-6x coverage, suggesting homoeologs are assembled separately. Extensive 454 sequence from normalized cDNA libraries has also been generated: 100,920 contigs (n50 886bp, longest contig 9873bp) totaling 909 Mb have been assembled to date. This is being supplemented with extensive Illumina cDNA sequence to generate a comprehensive transcriptome resource for bread wheat. This work will be completed by mid 2011. These assemblies are publically available at Cereals DB and form the first comprehensive genomic resource for the wheat research community world- wide. However, this resource is necessarily incomplete, unordered, and not yet highly accurate. In a complementary approach (supported by a BBSRC T&R grant (BB/G024650/1) and the EC Triticeae genome project) that aims to assess the potential of Illumina sequencing and assembly in wheat, we have assembled Illumina sequence reads from a purified wheat chromosome arm. Sequence and assembly methods have been optimized to create contigs with n50 of 775bp and max size 15kb covering 40% of the chromosome arm. Currently Illumina assemblies give an essentially complete coverage of genes identified by alignment to wheat cDNA sequence and syntenic Brachypodium genes, but only 25% of genes/cDNA are covered by a single Illumina assembly; the rest are covered by multiple contigs. Currently algorithms are being adapted and optimized to achieve assemblies with n50s closer to those achieved in simulations using sequenced wheat BACs (approximately 10kb). The main emphasis is on developing the memory-efficient assembly algorithms for building assemblies to enable larger-scale assembly. Another focus is on incorporating paired end read information directly into assemblies using novel computational approaches that also use memory most efficiently. The experience gained with this project is being applied to the IWGSC-coordinated Illumina survey sequencing of the 42 chromosome arms of Chinese Spring That began with the production of survey sequences to accompany the chromosome based physical mapping projects. Dr Rogers is coordinating this effort on behalf of the IWGSC and TGAC is contributing significantly with the production of 20-30-fold coverage in Illumina paired end sequences from 600bp fragments for 40% of the genome to generate sequence tags that will ascribe unambiguous homoeologous relationships to gene assemblies derived from whole genome sequencing.

The resources, knowledge and skills we have developed enable us to propose a feasible and cost-effective approach to generating a complete and accurate sequence of a majority of all wheat genes within the international collaborative framework provided by IWGSC. The research programme will be carried out in three inter-dependent themes. These aim to address the needs of the research community and industrial users in a timely, effective and efficient way.

The **reference genome sequencing theme** will work within the framework of IWGSC to contribute significantly towards a genome sequence with high utility for bread wheat Chinese Spring. This will be conducted in two phases: the first aims to deliver the accurate sequence of most bread wheat genes anchored within a comparative genomics framework, a first-pass annotation of genes, and database provision as an Ensembl genome browser. The second longer-term phase aims to anchor these genes in an accurate framework established by physical mapping and long-range sequence assemblies.

The **genome variation theme** aims to develop high throughput re-sequencing methods to access natural genetic variation in genes of selected wild wheat relatives and in ancestral species. This theme will be closely linked to the Wheat Pre-Breeding programme (WISP) to access genetic variation in key germplasm relevant to wheat pre-breeding applications. We also aim to access induced genetic variation in TILLING populations of tetraploid and hexaploid wheat to enable a wider range and depth of research in wheat. Work in this area will directly build upon current work in a BBSRC BBR grant (BB/I000712; Uauy- JIC and Philips- RRes) that is developing genome capture technology for re-sequencing and apply it on a whole genome scale.

The **database and outreach theme** aims to develop resources and expertise that enable rapid and effective uptake of the reference genome sequence and sequence variation. This involves maintaining and curating the sequence to facilitate longer-term use past the period of the grant, refining gene structures and their genomic positions, and updating gene annotations. We aim to conduct a series of specialist training and outreach events to promote wide uptake and use of the wheat genome sequence in the academic and relevant industrial research communities.

Research Programme

Theme 1. Creating a reference sequence of *Triticum aestivum* Chinese Spring 42.

Overall aim: to contribute to creating a high quality reference genome sequence of Chinese Spring bread wheat by 2015 within the framework of the International Wheat Genome Sequencing Consortium.
Theme Lead: TGAC, Partners: JIC; EBI.
Collaborators: IWGSC members: MIPS/IBIS Munich; IEB Olomouc and INRA Clermont-Ferrand; and CSHL Sequencing Centre.

Objective 1.1 Generate a highly representative wheat gene set from assemblies of whole genome and chromosome arm shotgun sequence reads.

We will build on the current 454 sequence assemblies (5.1x) and SOLiD sequence of the reference line Chinese Spring 42 (currently 20x) by sequencing whole genome DNA with Illumina (up to 40x coverage) to facilitate the integration with Illumina sequence data generated by the IWGSC for each of the chromosome arms (20-40x depth). 3kb and possibly 10kb. mate pair libraries will be used to optimize the potential for assembling longer contigs covering genes and capturing some of the adjacent sequence. The 5m 454 contigs are 5-6x depth and currently have an n50 of approximately 1 kb. Current analyses suggest these contigs have alignments to at least 75% of Brachypodium genes, demonstrating that they may be suitable templates for nucleating the assembly of Illumina reads using alignment methods such as MAQ and BWA to create a comprehensive non-redundant gene set. The sequence accuracy obtained from deep sequence coverage will identify sequence variation in large gene families necessary for distinguishing gene family members, including differences between homoeologs. The alignment will also permit correction of homopolymer and other sequence errors arising from low coverage in 454 data. In parallel, and in collaboration with the IWGSC and the individual chromosome leaders, the Illumina whole genome sequence will be assembled *de novo* with data from individual chromosome arms to identify genes not present in 454 sequence contigs, aiming to identify another 10-20% of wheat genes. This approach is increasingly feasible due to progress made in employing memory efficient assembly methods such as Cortex, SGA, SOAPdenovo or ALLPATHSLG to assemble wheat chromosome sequences. Recent progress in *de novo* assembly of genes from Illumina sequence of wheat chromosome 3DL, plant genomes of up to 1 Gb (sugar beet, tomato, potato), good progress in barley (19), together with the draft of the 2.25 Gb panda genome with Illumina sequence only (20), suggests that combining 454, SOLiD and Illumina reads into alignments and assembly will be a productive and rapid approach that can serve as a foundation for a high quality, finished reference sequence.

Experimental Plan

<u>1.1.1. Template preparation (months 0-24)</u>. High quality Chinese Spring 42 (CS-42) nuclear DNA (obtained from a single seed descent line with known provenance) free of organelle contamination and suitable for genome sequencing has been isolated for previous projects. Mate pair libraries (600 bp, 3kb and possibly 10kb) will be prepared for the Illumina platform and sequenced to generate up to 40-fold sequence coverage of the genome. The use of methylation sensitive enzymes (eg HpaII, SnaB1, AclI) to enrich for sequence templates from non-repeat DNA will also be used. This will directly support the assembly of gene and low copy sequences. From months 12-24 two methods to create sequence tags separated by relatively long defined distances in the wheat genome will be assessed. These sequences can provide a scaffold for assembling genic assemblies from 1.1.3 and 1.1.4 below. This will help create a *de novo* assembly that does not depend on predicted gene order, as described in 1.5.3 below. The first approach will use sheared DNA of 10 and 20 kb and incorporate *Eco*P15I adapters into the template preparation. This enzyme cleaves 27 bp distant from its recognition site and when sequenced provides a short sequence tag for SOLiD sequencing. This method is now starting to be used more generally (21). A complementary (and as yet untried) approach, will explore the efficiency and precise size selection of phage lambda packaging to create 40 kb jumping libraries. Approximately 40 kb sheared genomic DNA will be blunt ended, the ends ligated to an adapter containing a 12 bp cos site, and the DNA packaged with lambda extract to form concatemers .Packaged virions will be purified and DNA extracted, heated, blunt ended and ligated to a second biotinylated adaptor containing a *Not*I restriction site, circularised and sheared. The *Not*I cohesive ends will promote circularization. After streptavidin purification the mate pairs will be processed for Illumina sequencing. Assembly methods will be adopted to incorporate long- range sequence tags and paired end information into larger-scale *de novo* chromosome sequence assemblies. These approaches will be tested as part of TGAC's research programme and, if successful, they could remove the need for pooled BAC sequencing.

1.1.2. Sequence generation (months 0-6). To reach 40x coverage with Illumina, 3-4 Hi-Seq runs in 100 base paired-end mode will be carried out, focusing on generating deep coverage of long gap-size libraries.
1.1.3. Alignment and assembly (months 6-18). The Illumina and SOLiD sequences will be mapped to CS-42 454 assemblies using BWA or MAQ. The assemblies will be assessed by comparison with the predicted genes of the closely- related Brachypodium genome, aiming to achieve complete coverage of 75% of these, including flanking sequences. The assemblies will also be assessed by comparing them with assembled transcriptome data, the genomes of other grasses (see below) and with the results generated within the IWGSC annotation group. A key milestone will be achieving n50 in the kb range, after which assemblies will be released (See Objective 3.1 below).
1.1.4. *De novo* gene assembly (months 6-18). Using memory efficient algorithms such as Cortex (22) recently used to assemble the wheat 454 genome sequence and wheat chromosome 3DL from Illumina sequence) we will attempt a *de novo* assembly of gene space incorporating the Illumina reads, and in a subsequent step use Curtain (23) to integrate these with 454 assemblies and SOLiD reads. This should associate genes with surrounding sequences and enable annotation of intron / exon boundaries and promoter sequences. We will mask repetitive sequences or remove highly represented *k*mers in some assembly runs and assess assemblies and memory performance. This should identify new genes not detected by low-coverage 454 reads and possibly provide some linking scaffolds for longer range assemblies. In 1.5.3 below these assemblies will be scaffolded into larger chromosome and genome-wide assemblies.

Deliverables

- Sequence assemblies representing a non-redundant, comprehesive hexaploid wheat gene set, including the sequence of flanking sequences including promoters.
- Linking scaffolds for larger-scale genome assembly

Objective 1.2. Gene modeling, annotation, pseudogene identification, and the definition of gene relationships.

The non-redundant gene set defined in Objective 1.1 is predicted to represent a significant proportion of hexaploid bread wheat genes, recently estimated to be between 100,000- 150,000 (5). Pseudogenes are known to be particularly prevalent in bread wheat due to polyploidy and the action of retroelements, and these can be difficult to distinguish from intact protein-coding genes and wheat-specific genes. Therefore gene modeling and annotation will be designed to identify variants that may contain fragments of repeat sequences, suffer truncations or other changes. Pseudogenes have important recognized functions in wheat (21) and their identification and preliminary analysis will promote research in the consequences of polyploidy and gene evolution. The approach taken follows that used successfully for other grass genomes (7), and will be carried out in collaboration with IWGSC, by annotation teams at TGAC, MIPS/IBIS and INRA Clermont-Ferrand. In the polyploid wheat genome each gene has the potential to be present in three copies derived from the A, B and D genomes. The degree of sequence divergence between these homoeologs is in the same range as variation between members of gene families present in each genome. Therefore homoeologs can be distinguished in the 454 assemblies according to sequence differences, but further information is required to allocate gene models to their correct A, B or D genome. This information will be derived from Illumina survey sequence of purified chromosome arms, which will be completed during 2011. Deep Illumina coverage of the AA, DD and AABB genomes by Dick McCombie at CSHL (a collaborator supported by the NSF) will also provide a complementary resource with sufficient depth and coverage to identify sequence variation distinguishing homoeologs. Each gene locus will be given a unique 5 digit identifier that includes the chromosome arm identity and syntenic order (Objective 1.3) e.g. *Trae3DL01234*. Genes with good evidence that do not have an order will be numbered with the suffix u for unordered. Identities can then be subsequently modified to reflect a more precise gene order once this is defined.

Experimental Plan

1.2.1. Protein-coding gene annotation (months 12-24). Sequence assemblies will be annotated as part of the IWGSC activities involving TGAC, MIPS and INRA Clermont-Ferrand. Protein coding gene models will be defined using evidence- based predictions that incorporate transcriptome assemblies, protein homology from grasses and Arabidopsis, and *ab initio* gene finders such as Fgenesh++, and splice site models derived from rice and Brachypodium. Splice sites and 5' and 3'UTR sequences will be predicted from transcript assemblies. Gene predictions will be classified into confidence classes defined by their similarity to plant proteins in a reference database. Those with low confidence will be identified and assessed for Illumina transcriptome coverage to validate the gene prediction, and assessed for any repeat content different from paralogs, truncation or other sequence differences that may identify potential pseudogenes.

Deliverables

- Precise gene models and defined gene relationships and predicted gene functions of the majority of wheat genes.
- Identification and preliminary analysis of pseudogenes.
- Initial definition of homoeologous relationships.

Objective 1.3 Establish an approximate order of gene sequence assemblies based on conserved syntenic gene relationships and genetic maps.

Conserved gene order in grasses provides a powerful framework for trait mapping and identifying candidate genes. The complete genome sequences of rice, sorghum and Brachypodium, representing three diverse grass sub-families, have been aligned and a set of conserved syntenic genes identified in the "genome zipper" (13). This includes over 50% of the sequenced genes from these three species. Although it has been proposed that genome expansion in the large Triticeae genomes may erode collinearity (10), recent progress in barley genomics (14, 25) has managed to align approximately 75% of barley gene sequences to the zipper. Even if evidence from synteny and genetic maps is inconsistent with the homoeolog annotation, the gene could be assigned to the correct chromosome arm location by PCR of aneuploid lines by users. This data will be placed in a genome browser (see Theme 3) and will permit high resolution QTL and gene mapping, accelerate candidate gene identification, and provide a framework for establishing the precise order of all genes in subsequent studies.

Experimental Plan
1.3.1. Syntenic build of the wheat genome (months 18-30). Syntenic relationships will first be defined by sequence comparison of the annotated wheat genes to the sequenced genomes of Brachypodium, rice and sorghum by our IWGSC partner at MIPS/IBIS, who have recently reported a syntenic genome construction for barley. Aligning the wheat zipper with the wheat and *Ae. taushchii* genetic maps (for the D genome), emerging physical maps (from Objective 1.4 and current IWGSC work) and deletion mapped sequences will validate longer-range anchoring and assess homoeolog assignments. These alignments will be compiled into a zipper for each chromosome arm. Genes with high evidence (that is, no evidence for a pseudogene) that do not fit into the syntenic build will assigned to an unmapped bin on the relevant chromosome arm until additional physical and/or genetic map information (Objective 1.4) or long range sequence scaffold data (Objective 1.5) become available.

Deliverables

- An initial gene order based on synteny with Brachypodium and other grass genomes, integration of conserved gene order, genetic and deletion maps and emerging physical maps, and definition of homoeologous relationships.

Objective 1.4 Physical maps and reference sequence of four wheat chromosomes.

Wheat genes are predominantly found in small "islands" of 1-4 genes separated by extensive tracts of nested retroelement repeats (7) that can span up to 100-200 kb. To date the only feasible method for spanning, and eventually sequencing, these long tracts of intervening DNA and ordering genes precisely, is to construct a physical map of large insert BACs from purified chromosomes (11). This is the main strategy adopted by the IWGSC. These contigs are anchored and ordered by genetic maps, syntenic gene content, and deletions. This work has been completed for flow-sorted chromosome 3B (12) and 3A, while other chromosomes, including 3DL at JIC, are in progress coordinated by IWGSC (26). A minimal tiling path of BACs representing 3B is currently being sequenced using 454 at Genoscope in France, as part of the 3BSeq project (37), led by INRA Clermont Ferrand. Essentially complete sequence contiguity across BACs has been achieved using this approach and currently 12% of the chromosome has been completed (C. Feuillet, *pers comm*.). The IWGSC has allocated chromosomes 2B, 2D, 4B and potentially one other (to be determined later according to progress) to JIC/TGAC for physical mapping and sequencing. In this Objective we will define a minimal tiling path of BACs from a physical map, sequence these BACs, and then use the sequence to precisely order BAC contigs according to gene content, alignment with genetic and deletion maps and conserved gene order. This analysis will lead to the definition of precise gene order in a complete chromosome. The extent to which accurate long range assemblies of repeats can be achieved is presently unclear, but longer Illumina sequence reads, plus scaffold sequence reads from 1.1.1 and 1.5.1 and 1.5.2 below, will provide approximate sequence distances between genes.

Experimental Plan
1.4.1. Constructing physical maps of chromosomes (Months 1-36). BAC libraries from purified chromosomes 2B, 2D, 4B and one other (to be determined according to progress by other partners) will be provided by our IWGSC collaboration with the Institute of Experimental Botany (Olumuc, Czech Republic). They will be fingerprinted using an Illumina sequence- based protocol that involves making individual DNA preps, pooled, digested with a set of restriction enzymes, ligated to bar-coded adaptors, and fragments sequenced with 50 - 60 base reads to provide approximately 20-60 tags per BAC. The fingerprint contig programme fpc has been adapted to incorporate sequence tags and will be used to build contigs. BAC contigs will be placed in a long-range framework by alignment with the synteny maps, sequenced genetic markers, genetic maps and defined chromosome deletions. These resources will provide a detailed and validated long range sequence-based framework for both defining the order of genes and for identifying BACs for possible future finishing. This approach will be integrated with other groups working in IWGSC to contribute to generating a high resolution sequence tagged physical map of the entire bread wheat genome.
1.4.2. Generating a draft sequence of 4 wheat chromosomes (Months 1-48). Individual BAC DNA preps will be sheared and ligated to bar-coded adaptors that permit deconvolution of sequence reads from pools of 200-300 BACs. Pools of BAC DNA will then be sequenced in a single lane of an Illumina sequencer. The assembly parameters will be optimized to maximize the assembly of each BAC from sequences derived from a single fragment size (e.g.600 - 800bp), but in some cases it may be necessary to complement these data with paired end sequences from 3kb mate pair libraries. Assemblies will be annotated as in 1.2.1 above to identify genes and to define the sequence distances between genes. At this stage we cannot predict how well repeats may assemble, but progress with assembly of 454 sequences from BACs on chromosome 3B indicates it should be possible using Illumina read lengths of 2x 150bp.

Deliverables

- Physical maps of four wheat chromosomes
- Draft sequence of four wheat chromosomes, comprising the order and approximate sequence distance between genes and pseudogenes on the chromosomes.

Objective 1.5 Novel approaches to accurate long-range ordering of unigenes and whole genome assembly.

The BAC-based approach to defining chromosome-arm scale sequence scaffolds described above is feasible and it is currently the best approach to achieve long range chromosomal organization of sequences. However it is time-consuming and expensive due to the requirement to manage and prepare DNA for approximately ten thousand clones per chromosome. It is also limited by the precision of BAC anchoring to chromosomes. Research in this objective aims to explore the potential of new sequencing technologies for the rapid and cost effective identification of the chromosomal location of most wheat genes based on long-range sequence scaffolds. These approaches avoid cloning steps that are the most expensive and time-consuming component of current wheat genomics. Such scaffolds may also facilitate sequencing of intergenic regions in the future. The new mate-pair read technologies for Illumina to developed in 1.1.1 above will be used as sequence scaffolds to integrate support physical mapping work in IWGSC, and for *de novo* chromosome and genome assembly. This identification of most wheat genes will contribute significantly to the international goal to obtain a high quality finished reference genome sequence that includes both genic and intergenic regions.

Experimental Plan
<u>1.5.1. Exploring single molecule sequencing for sequence assembly (months 12-48).</u> Single molecule DNA sequencing technology is now a commercially available technology, although much needs to be done by users to achieve effective operation. A few sequencing centres have purchased the Pacific Biosciences machine, including our NSF- funded collaborator Dick McCombie at CSHL Sequencing Centre and the Wellcome Trust Sanger Institute. We aim to work with these groups to develop single molecule sequencing of wheat, in particular achieving high throughput sequencing of long reads (>1kb) and "strobe" sequencing to tag long DNA molecules with sequence reads at precisely defined intervals. There are many uncertainties in this approach, as the technology is still being developed, but we anticipate that by year 3 of the project our work and that of others investigating the potential of Pacific Biosciences single molecule sequencing and the technologies of other vendors will have managed to generate data in useful amounts and that methods for incorporating the novel types of data into sequence assemblies will be advancing.
<u>1.5.2. Whole chromosome / genome assembly (months 36-60).</u> Memory efficient sequence assembly algorithms specifically adapted to incorporate defined scaffold length reads as paired end information (derived from 1.5.1, 1.5.2 and 1.1.4 above) will be applied to assemble sequenced genes into larger regions of chromosomes with precisely defined distances between genes. The bioinformatics groups at TGAC and EBI are working of novel assembly tools to integrate contigs generated by deep sequencing platforms such as Illumina and SOLiD with low coverage tags (pair ends or strobes sequences) with the aim to generate long scaffolds. These developments are particularly relevant to the wheat genome as they are key to link contigs over regions of complex repeats. The assembly of long- range scaffolds will depend on the success of new sequencing technologies and sequence template preparation methods. Therefore we cannot predict how well the chromosomes will be covered by assemblies at this stage.

Deliverables

- Application of single molecule sequencing to the wheat genome to create spaced sequence tags for whole chromosome and ultimately whole genome assembly.
- Possible new methods for creating jumping libraries for scaffolding sequence assembly.
- First pass whole genome assembly of bread wheat.

Theme 2. Unlocking genetic variation in the Triticeae
Overall aim: to re-sequence multiple Triticeae genomes and mutant populations to identify useful sequence variation for breeding and biology.
Theme lead: JIC, Partners: TGAC; RRes; BBSRC Wheat Pre-Breeding Programme.

The availability of an accurate wheat non-redundant gene set within the first year of the project provides an opportunity to access sequence variation in wheat genes at a whole genome scale. Sequence variation in genes is only part of the spectrum of expected genetic variation in wheat, but it is a key part and of most interest to breeders and biologists interested in gene function. We have recently re-sequenced genomic DNA from the modern wheat cultivars Avalon, Cadenza, Rialto and Savannah using SOLiD sequencing, but this approach is currently too expensive for routine application to the many genomes and populations that need to be sequenced for gene discovery and pre-breeding research. To reduce the cost of wheat genome re-sequencing so it can be routinely applied on a large scale, we aim to use the non-redundant gene set to define a set of long oligos for genome capture (27) and Illumina sequencing. This method is now in general use, for example in maize (28), and it is also widely used to access sequence variation in human (29). It is also scalable, so sequencing can be focused on specific gene families in multiple lines. The challenge is to design long oligos that are both free of any repeat sequences and span genes appropriately. Good progress is being made by the JIC and RRes partners in defining these parameters on 3,000 genes in a BBSRC BBR project (BB/I000712). Oligo design will be further refined in this project using a wider range of genes, and scaled up to include the predicted 35,000 genes in each genome. We aim to design oligo probes to capture all three homoeologs with one set of oligos. The method will be applied in two distinct projects in this programme, and we aim to continue this beyond the end of this programme as a cost recovery genomic service, together with appropriate bioinformatics support.

Objective 2.1 Apply genome capture and re-sequencing to access sequence variation in the Triticeae.

The primary purpose of wheat genomics is to aid wheat breeding, biotechnology and biology projects by providing access to accurate and complete sequences of wheat genes. To access sequence variation in genes underlying key agronomic traits, we will work closely with the BBSRC Wheat Pre-Breeding Progamme to re-sequence the genomes of wheat landrace accessions such as those in the Watkins collection, and diploid and tetraploid Triticeae used in new synthetic hexaploid production. We aim to reduce the cost of re-sequencing to approximately £10,000 per genome, including Illumina sequencing. Assuming a gene space of approximately 500 Mb (100,000 genes of average size 5 kb including flanking sequences) and 30x coverage, 15 Gb of sequence is required for full genome re-sequencing. This can be achieved on a single lane of an Illumina Hi-Seq, so one Hi-Seq 2000 machine run could re-sequence 16 genomes simultaneously. Funds for technology development and sequencing approximately 10 genomes are included in the project. The precise lines to be sequenced will be determined in consultation with researchers in the pre-breeding programme. These will include single seed descent Watkins landraces that have useful traits (defined in yr 2 of that project) and Paragon, the elite parent to which Watkins landraces are being crossed. The sequence will be used to directly identify and fix desired landrace introgressions in back-cross generations. Other lines for re-sequencing include *Ae. tauschii, T. turgidum* and *T. dicoccoides* accessions used as parents for new synthetic wheat lines and in the chromosome segment substitution populations. The precise lines to be re-sequenced again depend on the phenotypes identified in the wheat pre-breeding programme. The sequence will be analysed in a specific bioinformatics pipeline utilizing alignment methods, and provided in a format revealing sequence differences from the reference genome and other lines- for example using an Ensembl genome browser and AnnoJ (30). These and other bioinformatic resources are described below.

Experimental Plan.

<u>2.1.1. SureSelect genome capture (months 24-60).</u> The Agilent SureSelect method uses long 120mer oligonucleotides spanning gene sequences to capture sheared genomic DNA fragments, which are then purified and sequenced using standard Illumina protocols. This method is now widely used (available in kit form) and has been pioneered in wheat by Keith Edwards' group at Bristol. We will design oligos based on the sequence of predicted genes with the highest level of annotation support (derived from 1.2.1 above), and these sequences will be further analysed to remove any repeat sequences from the oligo design. We will assess the method using an initial set of 10,000 genes spanning the genome to provide dense landmarks for applications in QTL and gene mapping, and then scaled to include approximately 30,000 genes. High quality nuclear DNA will be made from selected lines, captured and sequenced. Sequence reads will be aligned to the Chinese Spring 42 reference genes and polymorphisms identified and annotated. Sequence coverage of approximately 30x ensures high specificity. The polymorphism database will be publicly available, so users can design markers for selecting lines of interest in breeding programmes.

Deliverables

- A robust genome capture platform for resequencing of wheat genes
- A sequence database resource for defining sequence variation in wild wheat and Triticeae genes for application in wheat pre-breeding and gene function analysis

Objective 2.2. Identification of sequence variation in wheat TILLING populations.

Providing access to induced and natural sequence variation in genes facilitates biological investigation and is a key foundation of research in Arabidopsis and rice. We aim to sequence the gene space in tetraploid and hexaploid wheat populations that have been mutagenised with EMS in order to catalog sequence variation in genes and provide the lines and sequence information to users studying gene function. The populations are each approximately 2,000 individual M_2 lines, and high quality DNA preparations have already been made.

We propose to examine 1,500 individuals for both the tetraploid and hexaploid populations as this will provide a probability of >90% of identifying a knockout in any homoeologue in wheat (this increases with the length of the gene). The tetraploid TILLING population will be most useful for basic research projects as complete null mutants can be generated in a single generation by crossing the A and B genome mutants. On the other hand, hexaploid wheat represents the species of interest for commercial UK breeders so therefore we propose to sequence individuals from this population as well.

Experimental Plan

2.1.2. Sequencing wheat TILLING populations (months 24-60). The technical approach is the same as in 2.1.1 above, but the number of individual lines will be much larger and the sequence coverage will be correspondingly lower. DNA from individual lines will be sheared, size selected and adapted individually with 6-plex bar-codes. DNA will be pooled, selected by oligo capture as described in 2.1.1 above, adapted for Illumina template preparation, and sequenced on a single lane of a HiSeq2000 to generate approximately 15 Gb sequence per lane. This is equivalent to ~17x per genome considering 100,000 genes of average size 1.5kb as we will only focus on coding sequence for this approach. This coverage is sufficient for accurate detection of sequence variation even in heterozygous state as 66% of mutations are expected to be. Overall this requires 5000 Gb to cover 2000 selected genomes to 5x, or 25 Illumina HiSeq2000 runs. This method will be initially tested on 10,000 well- annotated genes, and scaled up to include up to 30,000 protein coding gene models for the complete populations. We will also develop the bioinformatic tools to visualize the identified mutations at both the DNA and protein level and use available software (PARSESNP; SIFT) to predict amino acid changes and effects on protein function. Each mutation will be traced to the original mutant plant whose seeds will be available for order through the JIC Germplasm Resources Unit.

Deliverables

- A mutation survey across 30,000 protein coding genes for tetraploid and hexaploid wheat TILLING populations, providing >90% probability of identifying a knockout mutation for each homoeologue.
- A sequence database resource for visualizing mutations across wheat genes, their predicted effect on the encoded protein and a link to the mutant stock for ordering.

Objective 2.3 Facilitating the further development of wheat genome selection technology and data analysis into service provision.

Once genome selection has been reduced to practice by meeting the Objectives described above (specifically to reduce the cost to approximately £10,000 per genome), we plan to apply our collective expertise to assessing the demand for developing a service for wheat re-sequencing and data analysis. This could be developed as spin- out company or it may build on current genotyping activities of iDNA Genetics, which JIC has a stake in. Given the many uncertainties facing the development of a successful enterprise, this Objective is currently an aspiration. However, given the specialist skills required for data analysis, a reasonable business model could be built on providing such a service for plant breeding.

2.3.1. Business Plan Development (months 48-60)

Once the protocol of wheat genome selection has been reduced to practice for a substantial proportion of wheat genes and can be conducted efficiently and routinely, within a tightly controlled cost framework, we will assess demand from the global wheat research community for service provision together with the JIC and TGAC Business Development Groups and PBL. Market analysis suggests that the major biotech companies, which all have major wheat breeding programmes, should be targeted. We will also establish if we have freedom to operate at a commercial level. Service could be provided on a range of scales, from genome scale surveys of sequence variation to assessing sequence variation in gene families.

Deliverables

- A Business Plan for cost effective long- term delivery of sequence service and analysis past the period of the grant application to serve academia and industry.

Theme 3. Developing an informatics infrastructure for Triticeae genomes to serve scientists, breeders and industry

Overall aim: to provide analysed data in useful formats for users from academia and industry.
Theme lead: EBI, Partners: TGAC, JIC.

The genome sequence of Chinese Spring 42 and sequence variation identified by re-sequencing Triticeae genomes and mutant populations represents a veritable deluge of sequence information. It will be a major challenge to analyse this sequence, to represent it in forms useful for a variety of users, and to facilitate data mining for new applications. In this theme, research in two Objectives will provide sequence data to users and provide a toolbox for applying genome sequence variation to wheat improvement.

3.1.1 Development of Ensembl-based services for wheat genomic data (months 1-60)

The Ensembl software infrastructure has already been used for the analysis and dissemination of data from a wide variety of genomes, including those of plants (31). EBI has a standard service delivery infrastructure for Ensembl-based services which will be used to additionally serve the wheat data. The Ensembl wheat database will be rebuilt at a minimum frequency of once per year, a process involving (i) incorporating any updates to the reference assembly (ii) the re-prediction of gene models, using newly available data (produced by this proposal, but additional any relevant data generated by other international projects) including reference assemblies, RNA sequence from wheat or other closely related species, protein sequence homologues from other plant species; and improved algorithms for gene prediction (iii) propagation of stable identifiers between successive releases and (iv) calculation of new cross-references (v) the incorporation of new SNP calls (vi) the production of data warehouses providing optimized access for common (gene- and SNP-centric) query types. Additionally, up to three times a year, an updated comparative analysis will be performed with other plant genomes, using adaptations of the existing DNA- (32) and protein- (33) centric pipelines developed at the EBI. Public access to data will be provided through an interactive browser, FTP, public mysql server, Perl API and query-optimised data warehouse. Between releases, we will work developing the Ensembl infrastructure for improved representation of polyploid genomes, to display relationships between genes within the three wheat genomes as well as between wheat and other species, and to make appropriate evolutionary inferences from the comparative evolutionary analyses.

Deliverables

- First release of Ensembl Wheat (month 24)
- Ensembl Wheat, first update (month 36)
- Ensembl Wheat, second update (month 48)
- Ensembl Wheat, third update (month 60)

3.1.2 Development of an infrastructure for the management of variation data (months 36-60)

We will develop an infrastructure for the integration and management of polymorphism data, using the Ensembl variation infrastructure (34), as previously utilised for other plants including Arabidopsis, rice, and grape. An interface will be developed that will distinguish between homeologous and heterologous SNPs, and to link out to phenotypic data and the results of QTL analysis. A SNP-centric data warehouse will be provided to allow users to find SNP sets matching search criteria (e.g. linkage to traits, association with genes, etc.). To provide a comprehensive resource, variation data from other wheat resequencing projects will be integrated with the data generated from this work. A facility will also be provided whereby users can predict the effect of additional (user-supplied) polymorphisms defined against the reference sequence. The database will be updated at intervals incorporating new sequence data as it becomes available and re-mapping existing data to the emerging reference.

Deliverables

- First release of Ensembl Wheat variation database (month 36)
- Variation database, first update (month 48)
- Variation database, second update (month 60)

Objective 3.2. Establish informatics tools to mine genome data.
In addition to the development of a long- term sustainable genomics resource described above, we aim to increase the impact for users by developing a new set of tools to analyse the wheat genome data for crop improvement. Critically, a resource for visualizing and mining sequence variation data and relating this to crop trait data, for example emerging from the pre-breeding programme. The resource will be derived from existing variation resources (34) as developed for other plants, but the scale of the wheat genome and its complexity will require improvements to speed, parallelization, resource consumption and the interface. In addition, a set of bioinformatics resources specifically targeted at breeding and gene discovery will be developed using the Galaxy system (35). In part these tools will be developed jointly with users according to their needs. Applications arising from these resources include mapping of sequence variation to genetically-defined intervals for QTL and gene identification using an adaptation of SHORE mapping (36), mapping sequence variation in biallelic mapping populations, and genome-wide association studies using sequence variation from multiple wheat lines. These require tools to mine sequence variation data (stored in a data warehouse at EBI, described above), linking this to gene functions, integration with field trial and other phenotyping data, and statistical analysis of sequence variation in populations. Data and analyses arising from use of this resource will be displayed and exported in standard formats to facilitate applications and to achieve a global impact.

3.2.1. Developing a wheat genome informatics toolbox (months 12-60).
Developing a database and query tools resource for relating genetic variation and crop traits
Developing a Galaxy resource for to mine this resource to exploit wheat genome sequence variation. Applications include genetic mapping in biallelic populations, QTL analysis, association genetics, gene expression analysis and chromatin analysis.

Deliverable

- An informatics infrastructure for applying wheat genomics resources for crop improvement.

Objective 3.3. Data Management, Outreach and Training
The genome sequence data and analyses generated in this project, and collectively in the IWGSC, represent a particularly large and complex dataset. To ensure that this resource is used to its full potential for wheat crop improvement and addressing food security, we aim to offer training courses tailored to the needs of users at the JIC Teaching Lab and at the Welcome Trust Genome Campus, focusing on ensuring the effective uptake of wheat genome technology into the breeding and biotech industries. The training courses will also involve the interactive development of bioinformatics tools based on users needs. To promote the use of the sequence resources we will discuss our progress and technical issues at specialist sequence technology meetings, at PAG in San Diego, where IWGSC meets, and in the UK cereal meetings (MONOGRAM and WGIN).

Plan

3.3.1. Joint training workshops (months 36-60). We will organize four joint science and training workshops, spaced through the 5 year research programme. Each workshop will focus on ensuring maximum impact from the main deliverables of the project by informing users of the resources, by training users in how to use it, by learning from potential users about their needs and how we can address these, and by promoting the work widely so other scientists can utilize it in their research. Workshops will be designed to give hands-on training and to promote engagement between specialist genomics scientists and users. To be effective in this format courses will be limited to approximately 25 participants and held in specialised training facilities at JIC/TGAC and EBI. A specific effort will be made to engage with and train scientists working in breeding companies or who wish to establish a career in plant breeding. Courses will be advertised widely to ensure we attract those whom we wish to reach, and we aim to ensure that UK industries needs are prioritized. The timing of workshops will also be adapted to progress in IWGSC to provide opportunities to maximize coordination.

Workshop 1 will be held in Month 24 during the construction of physical maps, sequencing of BACs and long-range scaffolding in Objectives 1.4 and 1.5. Its purpose is to impart the methods learnt and to review progress in integrating physical maps and sequence assemblies built from short read sequence data. The participants will be scientists working in wheat genomics and the meeting could form part of a larger IWGSC meeting. The main training component will focus on the wet lab work and on informatics for assembly.

Workshop 2 will be held in Month 30 when the wheat gene set has been assembled and annotated and the syntenic build of the wheat genome has been completed (See Deliverables from 1.1, 1.2 and 1.3). The course will focus on describing sequence assembly methods, gene identification and annotation, and how the syntenic build of the wheat genome was constructed. The participants will be mainly wheat biologists and breeders who wish to use the genome sequences. Training will be given in how to access the genome database via custom browsers and mine the data- for example to map genes and QTLs.

Workshop 3 will be held in Month 48 when genome selection methods are working optimally in Objectives 2.1 and 2.2. The purpose is to describe the methods and initial results and to engage potential users. A specific objective is to work jointly with the wheat Pre-Breeding programme to define lines to be resequenced, and the scope of resequencing. This workshop would also be a good opportunity to engage with UK plant scientists to learn more about the prospects of wheat genomics and to encourage them to include aspects of wheat biology in their research programmes.

Workshop 4 will be held in the final year of the project (Month 56 for example), at which time the wheat reference genome sequence would be nearing completion as part of IWGSC activities. The workshop will cover all areas of the project and scope out areas for future research in the application of wheat genomics to crop improvement. Depending on interest this could evolve into a larger meeting with external invited speakers. A good format for this is at the Plant and Animal Genome Conference, where there is a ready-made audience and speakers.

Project Management

The project will be managed by the PIs (Rogers, Caccamo, Bevan, Uauy, Philips, Kersey) who will meet quarterly via video conferencing to review progress and address any issues arising. Progress will also be discussed at regular monthly meetings on site in Norwich between TGAC and JIC. Web-based systems accessible to all partners and the Steering Committee will be used throughout the project to track progress and generate regular reports.

We will assemble a Steering Committee to advise, guide and support the project. It will be comprised of representatives of major stakeholders, the PIs and other experts in genome analysis. The stakeholders in this project include: The International Wheat Genome Sequencing Consortium, who provide the coordinating framework for the wheat genome project; a representative of the BBSRC, the main sponsors and link to the Global Food Security Programme (35); two members of the wheat breeding industry serving for 2 years on rotation to give each major breeding company exposure to the project; a scientist working in the human 1000 genome project to guide concept and technology development; and a scientist from the BBSRC Wheat pre-breeding Programme to provide a link to the application of genome data to pre-breeding. The

Steering Committee would meet the PIs annually and provide a progress report to BBSRC. Key management activities include establishing links with other genome variation projects, for the example the human and Arabidopsis 1000 genome projects, and rice and maize sequence variation projects, to develop jointly and use state of the art approaches and resources for mining genome sequence variation and applying this knowledge to crop improvement.

References

1. Birney, E. *et al* Nature 461, 168 (2009)
2. WASDE Report 488. November 2010. www.usda.gov/oce/commodity/wasde/
3. FAO: World Agriculture: towards 2030/2050. Interim Report
4. Feuillet, C., Leach, J.E., Rogers, J., Schnable, P.S. and Eversole, K. (2010) Trends in Plant Sciences, in press
5. Nature 436, 793 (2005); Science 326, 1112 (2010)
6. Nature Genetics 42, 961 (2010); Nature Genetics 42, 1027 (2010)
7. Choulet *et al* Plant Cell 22, 1686 (2010)
8. Levy & Feldman Plant Physiol. 130, 1587 (2002)
9. International Brachypodium Initiative Nature 463, 763 (2010)
10. Luo *et al* PNAS 106, 15780 (2009)
11. Dolezel *et al* Chromosome Res 15, 51 (2007); Bevan *unpublished*
12. Paux *et al* Science 322, 101 (2008)
13. Mayer et al Plant Physiol 151, 496 (2009)
14. K Mayer *pers comm*.
15. Eid et al Science 323, 133 (2008)
16. Handry et al Mol. Biol. Evol. 24, 1056 (2009)
17. Kushnir & Halloran Genetics 99, 495 (1981)
18. Uauy et al BMC Plant Biol. 9,115 (2009)
19. From presentations at the 10th Gatersleben meeting November 2010.
20. Li et al Nature 463, 311 (2010)
21. Young et al Genome Res 20, 249 (2010)
22. Caccamo et al ms submitted
23. Birney et al ms submitted
24. Griffiths et al Nature 439, 749 (2006)
25. N Stein, *pers comm*.
26. www.wheatgenome.org
27. Hodges et al Nature Genetics 39, 1522 (2007)
28. P. Schnable, pers comm.
29. Ng et al Nature 461, 272 (2009)
30. http://www.annoj.org/
31. Nucleic Acids Res. 2010 38:D563-9
32. Genome Res. 2008; 18: 1814–28
33. *G*enome Res. *2009;* 19*:* 327-35
34. BMC Genomics 2010, 11:293
35. Goecks et al Genome Biology 11, R86 (2010); http://main.g2.bx.psu.edu/
36. http://www.foodsecurity.ac.uk/
37. http://urgi.versailles.inra.fr/index.php/urgi/Projects/3BSeq