# Assembly of a Complex Genome: Defining Elements of Structure and Function

**James Breen**

**A Thesis presented for the degree of Doctor of Philosophy**

Murdoch
UNIVERSITY

**Murdoch University, Western Australia**

**September 2009**

**This thesis was supported by**

**Molecular Plant Breeding Co-operative Research Centre**

**(MPBCRC) and the Centre for Comparative Genomics (CCG),**

**Murdoch University**

# Abstract

The post-human genome sequencing project era has seen an influx of genome sequencing projects established to investigate the structure, composition and characteristics of plant genomes. While the genome sequences of smaller plant genomes (ie. Rice) are currently available, there has been a lack of progress on the study of large, complex genomes such as barley (*Hordeum vulgare*) and wheat (*Triticum aestivum*), due to the difficulties in their sequencing and assembly. The aim of this study is to assemble and annotate targeted regions of chromosome 3B from *Triticum aestivum* cv. Chinese Spring (CS) and Hope. This study also aimed to complete a comprehensive, inter- and intra-species comparative analysis using Bioinformatics tools and strategies, in order to define structural and functional elements within the genome

Genome sequences totalling 2.7Mb from two different loci of chromosome 3B in two different cultivars (*ctg11* from the short arm of CS, *ctg1034* from the long arm of CS and three assembled sequences over the equivalent *ctg11* region of Hope) were assembled using a novel 'two-phase' process that integrated information from a genome sequence assembler and a Triticeae-specific transposable element database. Through comparative genomics analysis a gene island was identified within a highly repetitive, heterochromatic region on 3BL that was highly conserved over four other cereal genomes (*Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor* and *Zea mays*). Chromodomain-containing long terminal repeats from the *gypsy* family of retrotransposons were identified adjacent to the gene island and may suggest an involvement in the targeted insertion of transposable elements at the loci, protecting the gene-island from

dynamic evolutionary change. Characterisation of the *ctg11* (*Sr2* region) genome sequence on 3BS, identified a large ~60kb mitochondrial genome insert and three members of the multi-gene beta-expansin family, with sequence analysis indicating local duplication within the sequence and rearrangements when compared to the equivalent region in a different wheat cultivar. *In silico* and real-time transcription analysis of the individual gene was also confirmed. Within the equivalent *ctg11* in Hope, a germin-like protein (GLP) cluster was identified and characterised that distinguishes between the two wheat cultivars. The genes in this GLP cluster were identified to belong to a sub-gene family that conferred broad level basal resistance in transient over-expressed systems in rice and barley.

The main outcome of this study was the development of a novel strategy of genome sequence assembly by utilising the complex component of the wheat genome that made assembly difficult: transposable elements. The complex genome sequence assembly methodology outlined in this thesis is suitable to be used as a model for future sequence assembly studies. The assembly of large pseudomolecule sequences (among the largest and most complete ever assembled in the wheat genome) enabled the Bioinformatics analysis of a representative sample of wheat chromosome 3B, providing valuable *in silico* outputs for future functional analyses and allowing an in-depth intra- and inter-species comparative analysis with related genomes.

# Declaration

Except where otherwise indicated, all work in this thesis is based on work carried out at the Centre for Comparative Genomics (CCG), Murdoch University, Australia. I declare that this thesis is my own account of my research and contains as its main content work, which has not previously been submitted for a degree at any tertiary education institution.

.....

James Breen

# Acknowledgements

There have been many people that have helped me throughout my PhD candidature. Firstly, I would like to thank all staff and students (past and present) from the Centre for Comparative Genomics, Murdoch University for their technical support and friendship over the last three years. Special thanks must go to Paula Moolhuijzen, Karon Ryan, Zayed Albertyn, Yair Motro and David Dunn for their extra support in Bioinformatics analysis. Thanks to Molecular Plant Breeding Co-operative Research Centre (MPBCRC) for their funding and professional development. I also thank all collaborators from France, USA, China and Switzerland. I would like to give special thanks to Thomas Wicker for his technical support and for hosting me in Zurich in January 2009.

My supervisors Professor Matthew Bellgard and Professor Rudi Appels deserve a lot of credit for this thesis. I thank Rudi for his enthusiasm, persistence, gentle encouragement and helpful advice, not only in this project, but also in my pursuit of future employment in genome research. I thank Matthew for taking a chance on me as an undergraduate student four years ago and having faith in my research.

Personally I would like to thank my parents, in-laws, grandparents, friends and extended family for their constant support. While they may not have understood much of my work, they were always excited and proud whenever I spoke about interesting developments. Special thanks also goes to my grandfather Professor Valentine Pervan for inspiring me to achieve excellence in academia. The high standards that he has set throughout his life is what I strive to reach.

Lastly I would like to thank my wife Nadine. Words cannot express the gratitude and appreciation I have for her support over the last three years. This thesis would not be possible without her. I dedicate this thesis to her.

# List of Abbreviations

| | |
|---|---|
| **BAC** | Bacterial Artificial Chromosome |
| **BLAST** | Basic Local Alignment Search Tool |
| **CCG** | Centre for Comparative Genomics |
| **CDD** | NCBI Conserved Domains Database |
| **CS** | Chinese Spring |
| **CSIRO** | Commonwealth Scientific and Industrial Research Organisation |
| **CSRDB** | Cereal smRNA Database |
| **DDBJ** | DNA Databank of Japan |
| **EBI** | European Bioinformatics Institute |
| **EMBL** | European Molecular Biology Laboratories |
| **EMBOSS** | European Molecular Biology Open Software Suite |
| **EST** | Expressed Sequence Tag |
| **FPC** | Fingerprinted Contig |
| **GBrowse** | Generic Genome Browser |
| **GFF** | Generic Feature Format |
| **GLP** | Germin-like Protein |
| **GRR** | Gene-Rich Region |
| **GyDB** | *Gypsy* Mobile Element Database |
| **HMM** | Hidden Markov Model |
| **InDel** | Insertion/Deletion |
| **ISBP** | Insertion-Site Based Polymorphism |

| | |
|---|---|
| **INRA** | Institut National de la Recherche Agronomique (France) |
| **IWGSC** | International Wheat Genome Sequencing Consortium |
| **LINE** | Long Interspersed Nuclear Element |
| **LTR** | Long Terminal Repeat |
| **MITE** | Transposable Element in Miniature |
| **MSU** | Michigan State University |
| **MTP** | Minimum Tiling Path |
| **MULE** | *Mutator*-like Transposable Element |
| **MYA** | Million Years Ago |
| **NCBI** | National Centre for Biotechnology Information |
| **NUMT** | Nuclear Mitochondrial Genomic Insert |
| **ORF** | Open Reading Frame |
| **OXOX** | Oxalate Oxidase |
| **PCR** | Polymerase Chain Reaction |
| **PDB** | Protein Databank |
| **PFAM** | Protein Family Database |
| **QTL** | Quantitative Trait Loci |
| **RT-PCR** | Real-Time Polymerase Chain Reaction |
| **SINE** | Short Interspersed Nuclear Element |
| **SOD** | Superoxidase Demutase |
| **SOE** | Son of Eric |
| **TE** | Transposable Element |

| | |
|---|---|
| **TIGR** | The Institute of Genomic Research |
| **TIR** | Terminal Inverted Repeat |
| **TREP** | Triticeae Repeat Element Database |
| **TSD** | Target Site Duplication |
| **UTR** | Untranslated Region |
| **WA** | Western Australia |

# List of Figure Legends

## General Introduction and Chapter 1

**Figure 1.1:** The conventional genome sequence assembly and annotation methodology employed in this study (on the left hand side). On the right hand side of the figure are general thesis questions investigating whether or not this conventional methodology is applicable to complex genomes such as hexaploid wheat.

**Figure 1.2:** Timeline of key dates in wheat agriculture.

**Figure 1.3:** Fertile Crescent region of the present-day Middle East. This region is believed to be the birthplace of modern agriculture and the starting point of wheat domestication and cultivation (taken from Salamini et al 2005).

**Figure 1.4**: Evolutionary history of polyploid wheat species. The shaded circles represent the cultivated forms of wheat. The broken lines represent the other possible origins of the AB Genome during formation of the hexaploid (Sabot et al. 2005)

**Figure 1.5:** Structural characteristics of hexaploid wheat chromosome 3B. Indicated on the left hand side of the figure are regions within the chromosome. On the right hand side of the figure are the chromosome 3B deletion bins (Saintenac et al. 2009).

**Figure 1.6:** Chromosome 3B consensus map showing all identified QTLs on the chromosome. This figure was taken from the CMap (Youens-Clark et al. 2009) map viewer (http://ccg.murdoch.edu.au/cmap/ccg-live/).

**Figure 1.7:** The structures of two types of Non-LTR Retrotransposons, LINEs and SINEs (Schmidt 1999).

**Figure 1.8:** The 'inner circle' of cereal genomics, where segmental rearrangements have shaped the formation of cereal genome sequences from the ancestral genome. Triticeae (wheat/barley), maize, sorghum and rice species are circled from the outside in, with related segments shown in parallel (taken from Bolot et al. 2009).

**Figure 1.9:** Integrated physical map at the telomeric end of chromosome 3BS and colinearity with rice. The figure shows the comparison of the wheat chromosome 3B genetic map (A), physical map contigs with their relative sizes (B) and its synteny with rice chromosome 1 (C) (Taken from Paux et al. 2008).

## Chapter 2

**Figure 2.1:** Two-phase genome sequence assembly and pseudomolecule construction methodology developed within this thesis chapter.

**Figure 2.2:** 'Phase 1' genome sequence assembly by utilising shotgun reads for contig ordering and rearrangement.

**Figure 2.3:** 'Phase 2' ordering of assembled BAC contigs into 'pseudomolecule' sequences using TEs and their element-specific target site duplications (TSDs).

**Figure 2.4:** Graph showing the relative positions of all 16 Chinese Spring wheat BAC clones on the 1.3Mb *ctg11* pseudomolecule sequence.

**Figure 2.5:** Pair-wise sequence comparison of the initial assembly of *ctg11* compared to the assembled BAC 036-I14 compiled from 454-technology sequencing.

**Figure 2.6:** Comparison of the two assembled 036-I14 BAC sequences from different sequencing technology methods (Sanger and 454-technology). *Ctg38* is annotated on the figure spanning the 3' inversion boundary.

**Figure 2.7:** Gel electrophoresis containing the 18F+21R producing a 5kb PCR product confirming the 5' junction of the of the *ctg11* inversion.

**Figure 2.8:** Genome Sequence annotation of *ctg1034* showing the 15 wheat BAC clones that were used in the assembly of the genomic sequence.

**Figure 2.9:** Gap closure of a genomic sequence gap in Hope contig 1. Primer sequences designed on either side of the gap were sequenced using BAC DNA and the product was compared to the genomic sequence in the top example. If the sequence product was identified at either side of the gap, the sequence was edited and closed (bottom example).

**Figure 2.10:** Sequencing trace file output from a PCR product produced in the gap closure of *ctg11* using 4Peaks (http://mekentosj.com/science/4peaks/). Large stretches of guanine (G) bases (indicated on the figure) cause a premature end to the sequence.

## Chapter 3

**Figure 3.1:** Location of *ctg1034*. The bin 3BL-7 is characterised by the SSRs Xgwm299, Xgwm2152, Xgwm547, Xgwm247, Xgwm340, Xgwm181, Xfwm4, Xcfa2170, Xbarc84, Xpsr170, XksuG62 (Sourdille et al. 2004) and the two ISBP markers (Sc3-119 and Sc3-120) could be assigned to this deletion bin. Lane 1-8 on the electrophoresis gel on the left-hand panel of the figure indicates the analysis of Sc3-119 and Sc3-120 with the deletion bins of chromosome 3B (lane 1: 3BS-8, lane 2: 3BS-9, lane 3: 3BS-1, lane 4: 3BL-10, lane 5: 3BL-7, lane 6: Halberd and lane 7: Cranbrook). Genetic mapping using the Cranbrook x Halberd population (McFadden et al. 2007) confirmed the long arm location on 3B. The genetic map (middle panel) shows a small part of the map (complete map, Cran*Hal 3B Feb09, is available at http://ccg.murdoch.edu.au/cmap/ccg-live/).

**Figure 3.2:** Predicted protein domain structure from the C-terminal integrase chromodomain

found in the *gypsy* LTR retrotransposable element named *Latidu*. The structure was predicted using sequence alignment with the characterised CLR4 chromodomain (1G6Z; Horita et al. 2001) from RCSB Protein Data Bank (Kouranov et al. 2006).

**Figure 3.3:** The 'classical' protein structure of two characterised chromodomains are shown in (A) along with the multiple sequence alignments of four families of chromodomains (B); chromodomains, shadow chromodomains, group I and II LTR retrotransposons chromodomains; and how they also contain 3 beta-sheets and one alpha-helical secondary protein structures (taken from Novikova 2009).

**Figure 3.4:** Multiple amino acid sequence alignments of the three syntenic and colinear gene island genes; *TaEP1* (A), *TaCRP1* (B) and *TaZFN1*(C). Each of the wheat, *Brachypodium* and maize genes were annotated in this study, while the *Sorghum bicolor* (v1.0 annotations from http://www.phytozome.net/sorghum) and rice (Ouyang et al. 2007) annotations. Annotated on the *TaCRP1* alignment (B) is the signal peptide and cysteine-rich domain structures outlined in Silverstein et al. (2007). The Clustal colour format is used (http://ekhidna.biocenter.helsinki.fi/pfam2/clustal_colours).

**Figure 3.5:** Genome annotation figure of the maize chromosome 3 genomic BAC AC217295.3 showing all six genes and Transposable elements. Dot-matrix diagram comparing the Maize Chromosome 3 BAC AC217295.3 60-130kb (horizontal axis) with *Sorghum bicolor* Chromosome 3 (71051-71068kb) using the DOTTER dotplot program is also shown.

**Figure 3.6:** Pair-wise sequence comparison of the wheat and *Brachypodium* gene island regions against the wheat and rice gene island regions. The wheat genes (*TaEP1*, *TaCRP1* and *TaZFN1*) are annotated on the figure and the arrows indicate the proposed insertion-deletion (InDel) events resulting in sequence movement between the species. Wheat contained an ~500bp InDel, while *Brachypodium* and rice showed ~200bp and ~6-7kb InDels respectively.

**Figure 3.7:** Plot of the distribution of smRNAs from the cereal small RNA database (CSRDB). The bars on the figure indicate the number of smRNAs found along a 50kb sliding window.

**Figure 3.8:** A gene island sequence summary figure next to a deduced evolutionary tree from analysed genome sequences homologous to the wheat chromosome 3BL *ctg1034* gene island. The red triangles indicate proposed InDel events identified in this study (such as those identified in Figure 3.6). A pair-wise sequence comparison between the maize and *Sorghum* located beneath the sequences shows the InDels between the two sequences from the start of the cysteine-rich peptide gene (*CRP1*) to the end of the Zinc finger protein (*ZFN1*) region.

**Figure 3.9:** Conservation of sequence between the each of the five syntenic copies of *EP1* in wheat, *Brachypodium*, rice, *Sorghum* and maize (in that order from top to bottom) using the ACT comparative sequence viewer. The yellow coloured blocks are the conserved exons. The red colour is all other conserved sequence with the intense colour being for the highest identity match.

**Figure 3.10:** Conservation of sequence between the five syntenic copies of *CRP1* in wheat, *Brachypodium*, rice, *Sorghum* and maize (in that order from top to bottom) using the ACT comparative sequence viewer. The yellow coloured blocks are the conserved exons. The red colour is all other conserved sequence with the intense colour being for the highest identity match.

**Figure 3.11:** Conservation of sequence between the five syntenic copies of *ZFN1* in wheat, *Brachypodium*, rice, *Sorghum* and maize (in that order from top to bottom) using the ACT comparative sequence viewer. The yellow coloured blocks are the conserved exons. The red colour is all other conserved sequence with the intense colour being for the highest identity match.

**Figure 3.12:** MSU (formerly TIGR) rice genome annotation of the rice chromosome 1 genome sequence over the gene island region (Ouyang et al. 2007). Highlighted on the figure is the proposed 6-7kb rice InDel that was absent when compared to the wheat sequence (Figure 3.6).

## Chapter 4

**Figure 4.1:** Genomic location, bacterial artificial chromosome (BAC) clone map and sequence annotation of a 357kb sub-sequence of the hexaploid wheat (*Triticum aestivum* cv Chinese Spring) *ctg11* genomic sequence from chromosome 3B containing three beta-expansin genes. Nucleotide sequence comparison between the *TaEXPB11cs2* and *TaEXPB11wy* cDNA sequence

(isolated from the Wyuna cultivar; Weichel et al. 2006) are shown.

**Figure 4.2:** Dot-matrix plot of 4kb of the genomic sequence surrounding *TaEXPB11cs2* (horizontal axis) against 4kb of the two other beta-expansin genes (*TaEXPB11cs1* and *TaEXPB11*c3) found within the *ctg11* Chinese Spring assembled genome sequence. Annotated on the figure are the exon regions of each gene. The blue box indicates a small 269bp CACTA DNA transposon located near the *TaEXPB11cs3* gene fragment.

**Figure 4.3:** Nulli-tetra lines for the group 3 chromosomes in Chinese Spring genetic stock lines analysed for occurrence of *TaEXPB11cs1* (left panel) and *TaEXPB11cs2* (right panel). For the chromosome assignment of *TaEXPB11cs1*, PCR products assaying exon 1 with one of the primers located in the insertion that characterises *TaEXPB11cs1* (see Material and Methods) were analysed on 2% agarose gels and stained with SYBR Green. The molecular markers are a 100 bp ladder (Axygen) and the product size of just over 1000 bp was as expected from the genome sequence (1027 bp). For the chromosome assignment of *TaEXPB11cs2* (right panel), SNPs in exon 1 that differentiated the genes on chromosome 3A, 3B and 3D were assayed by direct sequencing of PCR products from primers that amplified a common section of this exon. All possible nulli-tetra combinations were assayed and the sequences compared to the respective sequence from the *TaEXPB11wy*, *TaEXPB11hp* and *TaEXPB11cs2* genes (top of figure) in order to assign SNPs to particular chromosomes in the nulli-tetra combinations.

**Figure 4.4:** Pair-wise sequence comparison of the *TaEXPB11* genomic sequences of Wheat

cultivar Chinese Spring and an example of the short 'survey' sequences (spelt-red) carried out in this study on different wheat cultivars and spelt species. The three exons of *TaEXPB11cs2* are indicated below the figure (pale blue boxes, exon 1, 2 and 3 from left to right).

**Figure 4.5:** Gel analysis show samples from the RT-PCR (2% agarose gel) analysis of the internal control GADPH (top panel) and the first exon of TaEXPB11cs2 (lower panel). Material was analysed 7, 10, 15, 20 and 25 days post anthesis and included maternal tissue from the developing grain (M), endosperm (En) and embryonic (E) tissues of the wheat variety 'Cranbrook'.

**Figure 4.6:** ClustalW Multiple sequence alignment of two full-length beta-expansin genes found in wheat genomic sequences of chromosome 3B (*TaEXPB11cs1* and *TaEXPB11cs2*) compared to the *TaEXPB11* cDNA, rice homolog OsEXPB7 and maize Zea M 1 using the ClustalX program (Thompson et al. 2002). The blue (domain 1) and red (domain 2) lines above the sequence indicate the different domains and the signature EXPB motifs (the 'HFD' and conserved cysteine residues) are indicated below the sequence. The Clustal colour format is used (http://ekhidna.biocenter.helsinki.fi/pfam2/clustal_colours).

**Figure 4.7:** Protein model of the *TaEXPB11cs2* gene located within the chromosome 3BS *ctg11* region. Box A containing the blue/green wide ribbon is the lipoprotein A (RlpA)-like double-psi beta-barrel family domain (Domain 1) and box B, with the thinner red/orange ribbon, is the Pollen_allerg_1 grass type-2 pollen allergen domain (Domain 2). The red box located on the

edge of Domain 2 is the location of the 9bp insertion found in the spelt wheat varieties with the sequence comparison located beneath the figure.

## Chapter 5

**Figure 5.1:** Comparison of the genetic and physical map of the *ctg11* region on chromosome 3BS of *Triticum aestivum* indicating the BAC clones from both Chinese Spring (A) and Hope (B) cultivars (R. Mago et al. *in preparation*).

**Figure 5.2:** Genome annotation figure of the three Hope assembled pseudomolecule contigs and their relative sequence lengths.

**Figure 5.3:** Pair-wise sequence analysis of CS *ctg11* (from 1-100 kb) on the horizontal axis against Hope contig1 (92,064 bp) on the vertical. On each axis the sequence annotation is shown.

**Figure 5.4:** Pair-wise sequence analysis of CS *ctg11* (from 100-450 kb) on the horizontal axis against Hope contig 2 (290,884 bp) on the vertical. On each axis the sequence annotation is shown.

**Figure 5.5:** Pair-wise sequence analysis of CS *ctg11* (from 500-750 kb) on the horizontal axis against Hope contig 3 (207,617 bp) on the vertical. On each axis the sequence annotation is shown.

**Figure 5.6:** Pair-wise sequence alignment of Hope contig2 (230-290kb) against itself compared to the sequence annotation at the top of the figure (annotation to scale). Marked on the figure are the three different conserved blocks, containing one or more GLPs, that show high similarity. Flanking the three conserved blocks are TEs such as *gypsy* (yellow) and *copia* (red) LTR retrotransposons and CACTA DNA transposons (in blue).

**Figure 5.7:** Multiple sequence alignment of the wheat germin, germin-like and oxalate oxidase proteins located within NCBI GenBank compared with the four full-length germin-like genes found in this study. ClustalX was used to align all nine wheat amino acid sequences. The Clustal colour format is used (http://ekhidna.biocenter.helsinki.fi/pfam2/clustal_colours).

**Figure 5.8:** Phylogenetic tree (using un-weighted pair-group averages or UPGA with 1000 bootstraps) containing the Hope GLP genes (TaGLP1-4) along with germin gene sequences from rice (Manosalva et al. 2008) and barley (Zimmermann et al. 2006). The germin subfamilies indicated on the figure are characterised in Druka et al. (2002).

**Figure 5.9:** Crystallised structure of the manganese-containing (six manganese ions defined by green spheres) complex of germin proteins (top structure in the figure) that has a homohexamer structure made up of a trimer of dimers (bottom molecule in the figure). Each colour indicates a different dimer of germin proteins that make up the complex. Figure is taken from the structural characterisation study in Woo et al. (2000).

**Figure 5.10:** Comparison of the germin protein monomer (PDB accession: 1FI2) from Woo et al. (2000) on the left with the TaGLP1 protein structure identified in this study on the right. Both molecules were visualised using iMol (http://www.pirx.com/iMol/).

**Figure 5.11:** Summarised genomic sequence comparison between the syntenic *ctg11* regions of the CS and Hope wheat cultivars. The three assembled Hope contigs are separated by sequence gaps (GAP1 and GAP2) that are of unknown sizes.

## Chapter 6

**Figure 6.1:** Enhanced figure of the standard genome sequence assembly and annotation methodology (right hand side) introduced within the general introduction (Figure 1.1), with more emphasis placed on pseudomolecule construction and sequence analysis. The red arrows indicate the use of TE annotation in sequence assembly ('phase 2' assembly of pseudomolecules). On the left hand side are the research themes covered in the thesis and the contributions to them. Assembly of pseudomolecules was detailed in *chapter 2* while *chapters 3, 4 and 5* detailed the genome sequence analysis and annotation.

# List of Table Legends

## Chapter 1

**Table 1.1:** Table showing the QTLs, their defining markers located on physical contigs located within the chromosome 3B physical map. 'NC' indicates that no contig was identified that contained the marker but was located on a 3B neighbour map of co-segregating markers that are anchored to contigs (Paux et al. 2008).

**Table 1.2:** New transposable element classification system designed in Wicker et al. (2007) for all eukaryotic TEs. This system is based on the original two-class system separating retrotransposons and DNA transposons (Finnegan, 1989).

**Table 1.3:** Characteristics and pathogenicity of the wheat rusts (adapted from Table 10.12 from Anderson and Garlinge et al 2000; Pictures taken from GrainGenes http://wheat.pw.usda.gov).

**Table 1.4:** Genome project database statistics taken from NCBI Entrez Genome Project (http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html) taken on 25/04/2009.

## Chapter 2

**Table 2.1:** BAC clones used to assemble the chromosome 3B ctg11 genomic sequence.

**Table 2.2:** CAP3 Genome assembler run's using the 454-technology sequencing data of *ctg11* BAC 036-I14 from BGI, China. Seven different assembly parameters within the program were tested.

**Table 2.3:** Primer sequences designed for the analysis of the 5' and 3' junctions of the 60kb inversion region identified in 036-I14 from 454-technology sequencing.

**Table 2.4:** Primer designed for the closure of sequence gaps in *ctg11*

**Table 2.5:** Primer sequences for gap closure re-run from Table 2.4.

**Table 2.6:** 15 BACs used in the assembly of the *ctg1034* pseudomolecule.

**Table 2.7:** First group of primers designed for the closure of gap sequences in ctg1034 (* Products showing poor quality sequence repeated in Table 2.9).

**Table 2.8:** Second group of primers designed for the closure of gap sequences in ctg1034 (* Products showing poor quality sequence repeated in Table 2.9).

**Table 2.9:** Repeated primer sequences for gap closure (see Table 2.7 and 2.8).

**Table 2.10:** First group of primers designed for closure of gaps located within Hope contig 1.

**Table 2.11:** Second group of primers designed for closure of gaps located within Hope contig 1.

**Table 2.12:** First group of primers designed for closure of gaps located within Hope contig 2.

**Table 2.13:** Second group of primers designed for closure of gaps located within Hope contig 2.

**Table 2.14:** Primers designed for closure of gaps located within Hope contig 3.

## Chapter 3

**Table 3.1:** TE Annotation of chromosome 3BL *ctg1034* pseudomolecule.

**Table 3.2:** LTR dating analysis of complete LTR retrotransposons located within *ctg1034*.

**Table 3.3:** Confirmation of the LTR dates of a nested insertion.

**Table 3.4:** Gene and EST analysis of the *ctg1034* 'gene-island' genes.

**Table 3.5:** Wheat and Rice UniGene EST expression profiles of the three genes located within the *ctg1034* gene island. Barley EST profiles are also shown on the basis of BLASTN homology searches against the NCBI EST dataset. The eleven tissues are shown on the horizontal table axis along with the total tissue-specific ESTs located within the pool.

**Table 3.6:** Characteristics of three wheat genes identified to be syntenic and colinear genome sequences to the Rice, *Brachypodium* and *Sorghum bicolor* genome sequences.

**Table 3.7:** Gene Annotation of Maize Chromosome 3 BAC AC217295.3.

## Chapter 4

**Table 4.1:** RT-PCR primers used for expression analysis of *TaEXPB11cs2*-domain 2 and a GAPDH control.

**Table 4.2:** Gene-coding sequence annotation of 357,000bp sub-sequence of *ctg11* Assembled genomic BAC sequence containing the three EXPB11 genes.

**Table 4.3:** Mitochondrial genes regions located in the NUMT segment within the *ctg11* sequence and the number of EST sequences found (Zhang et al. 2009 *Submitted*; *Appendix I*).

## Chapter 5

**Table 5.1:** Gene annotation of the assembled Hope contig 1.

**Table 5.2:** Gene annotation of the assembled Hope contig 2.

**Table 5.3:** Gene annotation of the assembled Hope contig 3.

**Table 5.4:** Gene-coding regions common between Hope and CS and their genomic locations in both cultivars.

## Chapter 6

**Table 6.1:** NCBI GenBank sequence database contents of the four major agricultural crops: wheat, barley, rice and maize (http://www.ncbi.nlm.nih.gov; 17/08/2009).

# List of Submitted Journal Articles

**James Breen**, Thomas Wicker, Xiuying Kong, Juncheng Zhang, Wujun Ma, Etienne Paux, Catherine Feuillet, Rudi Appels and Matthew Bellgard, 2010, "A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: diverse gene function and genomic structure maintained in a tightly linked block", *BMC Plant Biology, 10:98*

**James Breen**, Dora Li, David S. Dunn, Ferenc Békés, Xiuying Kong, Juncheng Zhang, Jizeng Jia, Thomas Wicker, Rohit Mago, Wujun Ma, Matthew Bellgard and Rudi Appels, 2010, "Wheat beta-expansin (EXPB11) genes: Identification of the expressed gene on chromosome 3BS carrying a pollen allergen domain", *BMC Plant Biology, 10:99*

Juncheng Zhang, Jizeng Jia, **James Breen**, Rudi Appels and Xiuying Kong, 2009, "Recent insertion of a 52kb mitochondrial DNA segment in the wheat lineage", Submitted to BMC Plant Biology 25th August, 2009 (*See Appendix I*)

Frédéric Choulet, Thomas Wicker, Camille Rustenholz, Etienne Paux, Jérome Salse, Philippe Leroy, Stéphane Schlub, Marie-Christine Le Paslier, Ghislaine Magdelenat, Catherine Gonthier, Arnaud Couloux, Hikmet Budak, **James Breen**, Michael Pumphrey, Sixin Liu, Xiuying Kong, Jizeng Jia, Marta Gut, Dominique Brunel, James A. Anderson, Bikram S. Gill, Rudi Appels, Beat Keller, and Catherine Feuillet. 2010, "Megabase Level Sequencing Reveals Contrasted Organization and Evolution Patterns of the Wheat Gene and Transposable Element Spaces", *Plant Cell 10.1105/tpc.110.074187*

# Table of Contents

# General Introduction

Wheat is an important world food crop contributing >60% of the total calories consumed in the world daily. The temperate environment that wheat is able to be grown in means that it takes up more area of land that any other agricultural crop (Gill et al. 2004). In order to feed an exponentially increasing world population, crop yields need to be improved and research must be carried out to protect crops from abiotic-stress (extreme environmental conditions such as drought) and biotic-stress (attack from fungi, parasites and viruses that cause plant diseases).

Genome sequencing, enabling the identification of important agronomic genes, is an important step in improving crop characteristics and yields. Research into plant species such as *Oryza sativa* (rice), *Sorghum bicolor* and *Arabidopsis thaliana* have all benefited from a fully sequenced genome and their small and compact genomes allowed them to be sequenced relatively quickly and easily. Sequencing the wheat genome on the other hand has been impeded by its extremely large genome size (~16Gb; 37 times larger than the entire rice genome sequence), hexaploid genome structure and its high complexity (due a high proportion of repetitive sequences caused by transposable elements (TEs)).