Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Genomorientierte Bioinformatik

The genomic repertoire of complex and polyploid cereal genomes

Manuel Spannagl

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. C. Schwechheimer
Prüfer der Dissertation:	1. Univ.-Prof. Dr. H.-W. Mewes 2. Univ.-Prof. Dr. H. Schoof (Rheinische Friedrich-Wilhelms-Universität Bonn)

Die Dissertation wurde am 13. April 2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 23. Juni 2015 angenommen.

Abstract

Cereals such as wheat and barley are of utmost importance for human diet and are grown almost worldwide. Their genome sequences and gene repertoires, however, remained largely uncharacterized so far, due to large genome sizes, high repeat content and complex genome structures. To overcome limitations involved with the assembly of next generation sequencing data in cereal genomes such as collapsing of homeologous gene copies, in this work novel analysis strategies were developed to access the gene content of wheat and barley and construct gene families across closely related model and crop plants.

For the allohexaploid bread wheat genome, a 5x whole genome shotgun sequence survey was obtained and reads were mapped onto a set of ˜20,000 orthologous group representatives constructed from clustered gene families from related grass model plants [1]. Stringent sub-assembly of those reads resulted in the identification of about 94,000 distinct wheat transcripts which were separated and classified into their subgenome origin based on sequence similarities to the putative progenitors/subgenome donors *Aegilops tauschii*, *Aegilops sharonensis* and *Triticum urartu*. For that, several machine learning methods were trained, applied and evaluated on a chromosome-sorted sequence dataset from wheat chromosome 1. Support Vector Machines showed best results for the separation of homeologous genes with high overall precision (>70%) on about 66% of the gene assemblies which could be classified with high probability. Analysis of gene families with expanded copy numbers in the wheat genome identified, among others, NB-ARC domain containing proteins, involved in defense response mechanisms, F-box genes as well as storage proteins. Based on comparisons to gene family sizes in reference grass genomes, a gene retention rate between 2.5:1 and 2.7:1 was determined for the homeologous genes in wheat after polyploidisation about 8,000 years ago. Gene loss appeared to be similarly distributed across all subgenomes, indicating no subgenome dominance on the genomic level. The

identification of hundreds of thousands of gene fragments and additional gene domains highlights the ongoing pseudogenisation and dynamic evolution in the genome of bread wheat. The resources created within this work will significantly assist genome-based breeding efforts and variation selection in bread wheat whereas the orthologous assembly strategy developed here provides an efficient and powerful way to access the gene contents of other complex, previously uncharacterized, polyploid genomes, not limited to plants.

For barley, whole genome shotgun sequences were generated for the *Bowman*, *Barke* and *Morex* varieties and integrated into a comprehensive physical and genetic map framework with which more than 75% of the physical map contigs could be anchored to genetic positions on the barley chromosomes [2]. Assisted by comprehensive fl-cDNA libraries and RNA sequence expression data, gene prediction was performed on a *Morex* genome assembly, resulting in 26,159 high-confidence genes with homology support in other plant reference genomes. In addition, ˜27,000 novel transcriptionally active regions (nTARs) were identified on the barley genome, of which 4,830 respectively 2,450 appeared to be conserved in the *Brachypodium* and rice genomes. Comparative analysis of gene families with closely related species revealed sugar-binding proteins, sugar transporters, NB-ARC domain proteins as well as (1,3)-β-glucan synthase genes, potentially involved in plant-pathogen interactions, to be overrepresented in the barley genome.

All data generated within the analyses of the complex wheat and barley genomes were made available from a dedicated Triticeae PGSB PlantsDB database instance, providing access to genome sequences, gene calls and tools and interfaces to assist grass comparative genomics approaches [3].

Zusammenfassung

Getreidepflanzen wie Weizen oder Gerste werden weltweit angebaut und sind für die menschliche Ernährung von grösster Bedeutung. Die Genomsequenzen und die darin kodierten Gene sind für viele Getreidearten jedoch nicht oder nur teilweise beschrieben. Dies lässt sich vor allem auf die teilweise immensen Genomgrössen, den hohen Anteil an repetitiven Sequenzen sowie auf komplexe Genomstrukturen zurückführen. Um die daraus resultierenden Schwierigkeiten bei der Assemblierung von "next-generation"-Genomsequenzierungsdaten bei Getreiden zu reduzieren bzw. zu vermeiden wurden im Rahmen dieser Arbeit neuartige Methoden und Konzepte entwickelt und angewandt mit dem Ziel, die Gesamtheit der Gene im Genom von Weizen und Gerste zu beschreiben und damit Genfamilien im Kontext anderer, nah verwandter Pflanzenarten zu rekonstruieren und zu analysieren.

Mit Hilfe der 454-Sequenziertechnologie hergestellte Rohsequenzen des Genoms von Brotweizen, bestehend aus drei verschiedenen Subgenomen (allohexaploid), wurden auf rund 20,000 orthologe Referenzproteinsequenzen von nah verwandten Arten aligniert [1]. Die alignierten Weizensequenzen wurden daraufhin individuell für jedes Referenzprotein einzeln mit stringenten Assemblierungsparametern zusammengefasst. Daraus resultierten etwa 94,000 verschiedene Weizentranskripte welche schliesslich mit Hilfe von Sequenzähnlichkeiten zu ihren angenommenen Vorgängern *Aegilops tauschii*, *Aegilops sharonensis* und *Triticum urartu* einem Subgenom zugeordnet werden konnten. Dazu wurden verschiedene Algorithmen aus dem Bereich des maschinellen Lernens trainiert, angewandt und auf einem Datensatz mit chromosomen-sortierten Sequenzen eines einzelnen Weizenchromosoms evaluiert. *Support Vector Machine* Algorithmen wiesen dabei bei insgesamt hoher Präzision (>70%) auf etwa 66% der Genassemblierungen die besten Ergebnisse auf. Genfamilien mit expandierter Anzahl an Genkopien in Weizen enthielten unter anderem NB-ARC Domänen Proteine, welche

in verschiedenen Mechanismen zur Abwehrreaktion in Pflanzen eine Rolle spielen, sowie F-box Gene und Speicherproteine. Mit Hilfe von Vergleichen zu den Grössen von Genfamilien in verwandten Referenzorganismen konnte eine Rate zwischen 2.5:1 und 2.7:1 für die Beibehaltung von homologen Genkopien in Weizen nach der Polyploidisierung vor etwa 8000 Jahren ermittelt werden wobei sich der Genverlust gleich verteilt über die Subgenome darstellte. Dies deutet darauf hin dass in Weizen zumindest auf genomischem Niveau keine Dominanz eines einzelnen Subgenoms vorliegt. Die Identifizierung hunderttausender zusätzlicher Gen-fragmente und -domänen unterstreicht die andauernde Pseudogenisierung und evolutionäre Dynamik des Weizengenoms.

Die mit dieser Arbeit geschaffenen Ressourcen werden wesentlich dazu beitragen die genom-orientierte Züchtung sowie die Auswahl von genetischer Variation in modernem Saatweizen zu ermöglichen und zu unterstützen. Die hier erstmals genomweit angewandte Strategie der Assemblierung mit Hilfe orthologer Referenzproteine zeigt einen sehr effizienten Weg auf um den Geninhalt komplexer, bisher nicht charakterisierter, polyploider Genome zu entschlüsseln. Dieser Ansatz ist dabei nicht beschränkt auf pflanzliche Genome sondern kann überall dort Anwendung finden wo Genomgrösse und komplexe Genetik eine direkte Sequenzierung und Assemblierung der Genomsequenz verhindern oder erschweren.

Für das Genom von Gerste wurden mit Hilfe des *whole genome shotgun* Verfahrens Sequenzen für die Gerstenkultivare *Bowman*, *Barke* und *Morex* erzeugt [2]. Diese wurden in eine Struktur aus physikalischen und genetischen Karten integriert, womit schliesslich rund 75% der Sequenzcontigs aus der physikalischen Karten einer genetischen Position auf den Gerstenchromosomen zugewiesen werden konnten. 26,159 Genmodelle konnten auf der Genomsequenz von *Morex* mit hoher Zuverlässigkeit vorhergesagt werden, unterstützt von einer umfangreichen fl-cDNA Bibliothek sowie RNA Expressionsdaten. Zusätzlich wurden rund 27,000 *novel transcriptionally active regions (nTARs)* im Gerstengenom identifiziert von denen 4,830 bzw. 2,450 in den Genomen von *Brachypodium* und Reis konserviert sind. Die vergleichende Analyse von Genfamilien in Gerste mit nah verwandten Spezies ergab dass Zucker-bindende Proteine, Zucker-Transporter, NB-ARC Domänenproteine sowie *(1,3)-β-glucan synthase* Gene, welche möglicherweise eine Rolle spielen bei Pflanzen-Pathogen-Interaktionen, im Genome von Gerste überrepräsentiert sind.

Alle im Rahmen dieser Arbeit an den komplexen Genomen von Weizen

und Gerste erzeugten Daten und Ergebnisse, wie z.B. Genomsequenzen und Genvorhersagen, wurden in einer speziellen Triticeae Teildatenbank von PGSB PlantsDB abgelegt [3] und sind von dort aus für die Nutzer abrufbar und mit Hilfe von verwandten Referenzgenomen und dafür entwickelten Tools für eigene Analysen verfügbar.

List of publications

The following publications in peer-reviewed journals are described in this thesis:

1. Brenchley R*, **Spannagl M***, Pfeifer M*, Barker GL*, D'Amore R*, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, Kay S, Waite D, Trick M, Bancroft I, Gu Y, Huo N, Luo MC, Sehgal S, Gill B, Kianian S, Anderson O, Kersey P, Dvorak J, McCombie WR, Hall A, Mayer KF, Edwards KJ, Bevan MW, Hall N. *Analysis of the bread wheat genome using whole-genome shotgun sequencing.* Nature. 2012 Nov 29;491(7426):705-10. doi: 10.1038/nature11650. *joint first authors

2. **International Barley Genome Sequencing Consortium**. *A physical, genetic and functional sequence assembly of the barley genome.* Nature. 2012 Nov 29;491(7426):711-6. doi: 10.1038/nature11543. Epub 2012 Oct 17.

3. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, **Spannagl M***. *MIPS PlantsDB: a database framework for comparative plant genome research.* Nucleic Acids Res. 2013 Jan;41(Database issue):D1144-51. doi: 10.1093/nar/gks1153. Epub 2012 Nov 29. *corresponding author

Additional publications by the author:

1. Chaki M, Kovacs I, **Spannagl M**, Lindermayr C. *Computational Prediction of Candidate Proteins for S-Nitrosylation in Arabidopsis thaliana.* PLoS One. 2014 Oct 21;9(10):e110232. doi: 10.1371/journal.pone.0110232.

2. **International Wheat Genome Sequencing Consortium (IWGSC)**. *A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome.* Science. 2014 Jul 18;345(6194):1251788. doi: 10.1126/science.1251788.

3. Marcussen T, Sandve SR, Heier L, **Spannagl M**, Pfeifer M, International Wheat Genome Sequencing Consortium, Jakobsen KS, Wulff BB, Steuernagel B, Mayer KF, Olsen OA. *Ancient hybridizations among the ancestral genomes of bread wheat.* Science. 2014 Jul 18;345(6194):1250092. doi: 10.1126/science.1250092.

4. Pfeifer M, Kugler KG, Sandve SR, Zhan B, Rudi H, Hvidsten TR, **International Wheat Genome Sequencing Consortium**, Mayer KF, Olsen OA. *Genome interplay in the grain transcriptome of hexaploid bread wheat.* Science. 2014 Jul 18;345(6194):1250091. doi: 10.1126/science.1250091.

5. Mathew LS*, **Spannagl M***, Al-Malki A, George B, Torres MF, Al-Dous EK, Al-Azwani EK, Hussein E, Mathew S, Mayer KF, Mohamoud YA, Suhre K, Malek JA. A first genetic map of date palm (Phoenix dactylifera) reveals long-range genome structure conservation in the palms. BMC Genomics. 2014 Apr 15;15:285. doi: 10.1186/1471-2164-15-285. *joint first authors

6. Kugler KG, Siegwart G, Nussbaumer T, Ametz C, **Spannagl M**, Steiner B, Lemmens M, Mayer KF, Buerstmayr H, Schweiger W. *Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (Triticum aestivum L.).* BMC Genomics. 2013 Oct 24;14:728.

doi: 10.1186/1471-2164-14-728.

7. **Spannagl M**, Martis MM, Pfeifer M, Nussbaumer T, Mayer KF. *Analysing complex Triticeae genomes - concepts and strategies.* Plant Methods. 2013 Sep 6;9(1):35. doi: 10.1186/1746-4811-9-35.

8. Silvar C, Perovic D, Nussbaumer T, **Spannagl M**, Usadel B, Casas A, Igartua E, Ordon F. *Towards positional isolation of three quantitative trait loci conferring resistance to powdery mildew in two Spanish barley landraces.* PLoS One. 2013 Jun 24;8(6):e67336. doi: 10.1371/journal.pone.0067336.

9. Munoz-Amatriain M, Eichten SR, Wicker T, Richmond TA, Mascher M, Steuernagel B, Scholz U, Ariyadasa R, **Spannagl M**, Nussbaumer T, Mayer KF, Taudien S, Platzer M, Jeddeloh JA, Springer NM, Muehlbauer GJ, Stein N. *Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome.* Genome Biol. 2013 Jun 12;14(6):R58. doi: 10.1186/gb-2013-14-6-r58.

10. Vigeland MD, **Spannagl M**, Asp T, Paina C, Rudi H, Rognli OA, Fjellheim S, Sandve SR. *Evidence for adaptive evolution of low-temperature stress response genes in a Pooideae grass ancestor.* New Phytol. 2013 Sep;199(4):1060-8. doi: 10.1111/nph.12337.

11. Jia J, Zhao S, Kong X, Li Y, Zhao G, He W, Appels R, Pfeifer M, Tao Y, Zhang X, Jing R, Zhang C, Ma Y, Gao L, Gao C, **Spannagl M**, Mayer KF, Li D, Pan S, Zheng F, Hu Q, Xia X, Li J, Liang Q, Chen J, Wicker T, Gou C, Kuang H, He G, Luo Y, Keller B, Xia Q, Lu P, Wang J, Zou H, Zhang R, Xu J, Gao J, Middleton C, Quan Z, Liu G, Wang J, International Wheat Genome Sequencing Consortium, Yang H, Liu X, He Z, Mao L, Wang J. *Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation.* Nature. 2013 Apr 4;496(7443):91-5. doi: 10.1038/nature12028.

12. Gaupels F, Sarioglu H, Beckmann M, Hause B, **Spannagl M**, Draper J, Lindermayr C, Durner J. *Deciphering systemic wound responses of the pumpkin extrafascicular phloem by metabolomics and stable isotope-coded protein labeling.* Plant Physiol. 2012 Dec;160(4):2285-99. doi: 10.1104/pp.112.205336.

13. **Tomato Genome Consortium**. *The tomato genome sequence provides insights into fleshy fruit evolution.* Nature. 2012 May 30;485(7400):635-41. doi: 10.1038/nature11119.

14. Fröhlich A, Gaupels F, Sarioglu H, Holzmeister C, **Spannagl M**, Durner J, Lindermayr C. *Looking deep inside: detection of low-abundance proteins in leaf extracts of Arabidopsis and phloem exudates of pumpkin.* Plant Physiol. 2012 Jul;159(3):902-14. doi: 10.1104/pp.112.198077.

15. Young ND, Debelle F, Oldroyd GE, Geurts R, Cannon SB, Udvardi MK, Benedito VA, Mayer KF, Gouzy J, Schoof H, Van de Peer Y, Proost S, Cook DR, Meyers BC, **Spannagl M**, Cheung F, De Mita S, Krishnakumar V, Gundlach H, Zhou S, Mudge J, Bharti AK, Murray JD, Naoumkina MA, Rosen B, Silverstein KA, Tang H, Rombauts S, Zhao PX, Zhou P, Barbe V, Bardou P, Bechner M, Bellec A, Berger A, Berges H, Bidwell S, Bisseling T, Choisne N, Couloux A, Denny R, Deshpande S, Dai X, Doyle JJ, Dudez AM, Farmer AD, Fouteau S, Franken C, Gibelin C, Gish J, Goldstein S, Gonzalez AJ, Green PJ, Hallab A, Hartog M, Hua A, Humphray SJ, Jeong DH, Jing Y, Jöcker A, Kenton SM, Kim DJ, Klee K, Lai H, Lang C, Lin S, Macmil SL, Magdelenat G, Matthews L, McCorrison J, Monaghan EL, Mun JH, Najar FZ, Nicholson C, Noirot C, O'Bleness M, Paule CR, Poulain J, Prion F, Qin B, Qu C, Retzel EF, Riddle C, Sallet E, Samain S, Samson N, Sanders I, Saurat O, Scarpelli C, Schiex T, Segurens B, Severin AJ, Sherrier DJ, Shi R, Sims S, Singer SR, Sinharoy S, Sterck L, Viollet A, Wang BB, Wang K, Wang M, Wang X, Warfsmann J, Weissenbach J, White DD, White JD, Wiley GB, Wincker P, Xing Y, Yang L, Yao Z, Ying F, Zhai J, Zhou L, Zuber A, Denarie J, Dixon RA, May GD, Schwartz DC, Rogers J, Quetier F, Town CD,

Roe BA. *The Medicago genome provides insight into the evolution of rhizobial symbioses.* Nature. 2011 Nov 16;480(7378):520-4. doi: 10.1038/nature10625.

16. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, **Spannagl M**, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL. *The Arabidopsis lyrata genome sequence and the basis of rapid genome size change.* Nat Genet. 2011 May;43(5):476-81. doi: 10.1038/ng.807.

17. Mewes HW, Ruepp A, Theis F, Rattei T, Walter M, Frishman D, Suhre K, **Spannagl M**, Mayer KF, Stümpflen V, Antonov A. *MIPS: curated databases and comprehensive secondary data resources in 2010.* Nucleic Acids Res. 2011 Jan;39(Database issue):D220-4. doi: 10.1093/nar/gkq1157.

18. **Spannagl M**, Mayer K, Durner J, Haberer G, Fröhlich A. *Exploring the genomes: from Arabidopsis to crops.* J Plant Physiol. 2011 Jan 1;168(1):3-8. doi: 10.1016/j.jplph.2010.07.008. Review.

19. **International Brachypodium Initiative**. *Genome sequencing and analysis of the model grass Brachypodium distachyon.* Nature. 2010 Feb 11;463(7282):763-8. doi: 10.1038/nature08747.

20. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, **Spannagl M**, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KF, Messing J, Rokhsar

DS. *The Sorghum bicolor genome and the diversification of grasses.* Nature. 2009 Jan 29;457(7229):551-6. doi: 10.1038/nature07723.

21. **Spannagl M**, Haberer G, Ernst R, Schoof H, Mayer KF. *MIPS plant genome information resources.* Methods Mol Biol. 2007;406:137-59.

22. Klee K, Ernst R, **Spannagl M**, Mayer KF. *Apollo2Go: a web service adapter for the Apollo genome viewer to enable distributed genome annotation.* BMC Bioinformatics. 2007 Aug 30;8:320.

23. **Spannagl M**, Noubibou O, Haase D, Yang L, Gundlach H, Hindemitt T, Klee K, Haberer G, Schoof H, Mayer KF. *MIPSPlantsDB–plant database resource for integrative and comparative plant genome research.* Nucleic Acids Res. 2007 Jan;35(Database issue):D834-40.

24. Haberer G, Mader MT, Kosarev P, **Spannagl M**, Yang L, Mayer KF. *Large-scale cis-element detection by analysis of correlated expression and sequence conservation between Arabidopsis and Brassica oleracea.* Plant Physiol. 2006 Dec;142(4):1589-602.

25. Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T, **Spannagl M**, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KF, Rogers J, Quetier F, Oldroyd GE, Debelle F, Cook DR, Retzel EF, Roe BA, Town CD, Tabata S, Van de Peer Y, Young ND. *Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes.* Proc Natl Acad Sci U S A. 2006 Oct 3;103(40):14959-64. Epub 2006 Sep 26. Erratum in: Proc Natl Acad Sci U S A. 2006 Nov 21;103(47):18026. Scheix, Thomas [corrected to Schiex, Thomas].

26. Schoof H, **Spannagl M**, Yang L, Ernst R, Gundlach H, Haase D, Haberer G, Mayer KF. *Munich information center for protein sequences plant genome resources: a framework for integrative and comparative analyses 1(W).* Plant Physiol. 2005 Jul;138(3):1301-9.

Acknowledgments

First of all I want to thank my supervisors Dr. Klaus Mayer and Prof. Dr. Hans-Werner Mewes. Klaus supported my career for more than 10 years now and encouraged me to write this thesis. Without his continuous advice and extremely helpful discussions this thesis could not have been completed in its current form. Thanks Klaus, for always having your door open for questions and problems and sharing your great knowledge and experience about science! Klaus also provided the possibility to work in a number of exciting and challenging projects as well as within a very cooperative group, both very important factors for the success of this thesis (and everyday work). Prof. Mewes kindly gave me the opportunity to write my PhD thesis in his department and provided valuable advice over the full course of this thesis.

I also want to thank Prof. Dr. Heiko Schoof who gave me the opportunity to join the MIPS plant group initially. Heiko shares his knowledge with great patience and extremely helped making my start into science easier.

A big thanks goes to all members of the MIPS/PGSB plant group who were always there to discuss things and help with problems or questions. I especially want to thank Matthias Pfeifer for the excellent collaboration in the UK wheat project as well as Thomas Nussbaumer, Dr. Heidrun Gundlach, Dr. Kai Bader and Mihaela Martis for working together with me in the barley sequencing project and/or on PlantsDB. Finally I want to thank Dr. Georg Haberer who supported my work with great discussions and priceless advice as well as Dr. Remy Bruggmann for ongoing encouragement.

This work would not have been possible without our cooperation partners and their reliance and willingness to share data and ideas. In the first place I want to thank all members of the UK wheat consortium as well as those from the IBSC (International Barley Sequencing Consortium). From the UK wheat group I especially want to acknowledge Rachel Brenchley for the great collaboration as well as Prof. Michael Bevan, Prof. Neil Hall, Prof.

Keith Edwards and Prof. Anthony Hall...it was a pleasure for me to be able to work with them. Thanks for excellent discussions and meetings. From the IBSC I especially want to thank our partners at IPK Gatersleben for the close collaboration and interaction, Dr. Nils Stein and Dr. Uwe Scholz in particular.

Last but not least I would like to thank my wife Christine for her loving support in all aspects of writing this thesis - from initial encouragement to discussions on the science on to very helpful advice with writing and finishing this thesis. And of course for giving me a motivating example on how to do a PHD thesis! Finally I want to thank my family which always supported my education and provided both retreat and encouragement.

Contents

List of Figures

List of abbreviations

454 454 Life sciences, `http://my454.com/`.

BAC Bacterial Artificial Chromosome

BBH Best Bidirectional Hit

Bp base pairs

CNV Copy Number Variation

EST Expressed Sequence Tag

flCDNA full length cDNA

Gbp Giga base pairs

GO Gene Onthology

IBSC International Barley Sequencing Consortium

IWGSC International Wheat Genome Sequencing Consortium

LCG low-copy-number genome assembly

Mbp Mega base pairs

MIPS Munich Information Center for Protein Sequences, `http://mips.helmholtz-muenchen.de/`

MTP Minimum Tiling Path

MYA Million years ago

NGS Next Generation Sequencing

nTAR novel transcriptionally active region

OG Orthologous Group

PGSB Plant Genome and Systems Biology, `http://pgsb.helmholtz-muenchen.de/plant/genomes.jsp`

SNP Single Nucleotide Polymorphism

WGD Whole Genome Duplication

WGS Whole Genome Shotgun

Chapter 1

Introduction

1.1 Focus and objectives of this study

Over the last couple of years, dozens of plant genomes have been sequenced, due to cost-efficient, high-throughput and fast next generation sequencing technologies [4-7]. The genome sequences of plants are an important resource for breeders, biologists and plant researchers for many reasons: the genome sequence and the genes encoded in it facilitate plant breeders to identify and select for specific traits related to e.g. yield, disease resistance and cold/drought tolerance [8]; the genome sequence enables biologists to search and identify genes responsible for specific phenotypes and genes involved in pathways under investigation [9]; genome sequences from multiple, related plants help to understand and study the complex evolution of plants [10, 11]; and finally, plant genome sequences provide a substantial basis to study natural variation within populations and relationships, differences and similarities among related plant species [12].

However, the genomes of many important cereals including bread wheat and barley bear great challenges for sequencing and analysis due to their large size, high repeat content (over ˜80%) and complex genomics. With 5.1 Giga-basepairs (Gbp) in size, the genome of barley is almost double the size of the human genome (˜3 Gbp). The barley genome is diploid (2n) with a total of 7 chromosomes. The genome of bread wheat has a total size of ˜17 Gbp and is composed of three different diploid subgenomes and is thus allohexaploid (6n). High sequence identity (˜97%) between the homeologous genes of the subgenomes complicate their assembly and separation and ask for novel analysis strategies and concepts. A more detailed introduction into the genome characteristics of cereals is given in chapter 1.3.

As a result, the genome repertoires of important crop plants such as wheat and barley remained largely uncharacterized until recently, with limited knowledge about gene content, gene family composition, pseudogenisation rates and other genetic elements. In this thesis a number of open questions related to the genome biology of Triticeae plants have been examined and new concepts for the analysis of large and complex plant genomes are proposed. For this, genome sequencing data for wheat and barley were used that were generated within the UK wheat consortium and the International Barley Sequencing Consortium (IBSC) (see 1.3.1 for more details on the sequencing data and sequencing consortia). Objectives in this study include:

- Analysis of the gene content in the complex and large genomes of the Triticeae wheat and barley including gene prediction, functional annotation and comparison to other plant genomes;
- Analysis of the gene family composition in the complex and large genomes of Triticeae including the identification of expanded and contracted gene families and their functional roles in Triticeae biology;
- Identification of novel transcribed regions (nTARs) in the genomes of Triticeae and analysis of their conservation in related species;
- Identification of species- , Triticeae- and grass-specific genes and gene families and the elucidation of their potential functional role and impact in/for Triticeae biology;
- Fate of homeologous genes in polyploid grass genomes such as bread wheat: is there any preferential gene loss in one of the subgenomes and if yes, to what degree? What is the overall gene retention rate after polyploidisation in the bread wheat genome? Are specific functional categories of genes/gene families more retained or faster evolving/degrading (pseudogenisation rate)? What is their functional role in the Triticeae? What level of divergence between homeologous wheat genes can be observed?
- New concepts for the analysis of complex Triticeae genomes: Reconstruction of homeologous genes in a polyploid genome from NGS shotgun data (short reads); Separation of homeologous genes (gene fragments) in a polyploid genome and classification of their subgenome origin;

- Integration, data management and visualisation of heterogenous and complex genome data from Triticeae genome sequencing and analysis projects within the PlantsDB database framework;

In the introductory part of this thesis I will first outline the characteristics and evolution of plant genomes in general (section 1.2), with a more detailed view on the pecularities and challenges involved with the analysis of the complex genomes of Triticeae (section 1.3). Here, I will also introduce the sequencing data and sequencing consortia which provided the foundation for the analyses described in this thesis (section 1.3.1). With an overview on the taxonomy and economic importance of Triticeae plants, section 1.4 emphasizes the relevance of this work for applications in plant biology and agriculture and provides background knowledge about phylogenomic relationships among Triticeae (relevant for comparative genomics approaches introduced later). In order to identify and analyse the gene content and gene families in Triticeae genomes, section 1.5 aims to introduce the objectives and targets as well as basic concepts and methods for the identification of conserved and species-specific gene models and the computation of gene families. Resulting from the novel methods developed and the genome analyses carried out in this study, heterogenous and complex Triticeae genome data had to be integrated from different resources and managed in a dedicated database framework as well as disseminated through specialized tools. Section 1.6 gives an introduction into existing genome database systems and outlines the specific needs for the integration and management of the data types generated also in this study. Section 1.6.1 finally describes ways and technologies to aggregate genome data from distributed genome resources and databases. This aspect becomes increasingly important when working with the bread wheat and barley genome data described in this thesis as no single data repository or database framework exists.

1.2 Evolution and characteristics of plant genomes

1.2.1 Plant genome sizes and variation

Within the plant kingdom, genome sizes show a high degree of variance. *Arabidopis thaliana* (thale cress) was the first plant to be fully sequenced in 2000 [13] not least because of its relative small genome size of about 125 Mega-basepairs (Mbp). Comparably medium-sized plant genomes are represented by e.g. rice (˜389 Mbp) [14], tomato (˜900 Mbp) [15], *Medicago*

truncatula (barrel medic, ~375 Mbp) [16], *Brachypodium distachyon* (purple false brome, ~272 Mbp) [17] and *Sorghum bicolor* (sweet sorghum, ~730 Mbp) [4]. Larger genome sizes are observed for maize (~2,300 Mbp) [18], barley (~5,100 Mbp) [2] and bread wheat (~17,100 Mbp) [1]. However, plants also contribute to some of the largest genomes known today, with ~149,000 Mbp [19] for *Paris japonica* and many more [20].

Figure 1.1 summarizes the genome sizes of some important plants and puts them into relation with the genomes of important non-plant species, such as bacteria (*E.coli*), yeast, fruit fly (*D. melanogaster*) and the human genome.

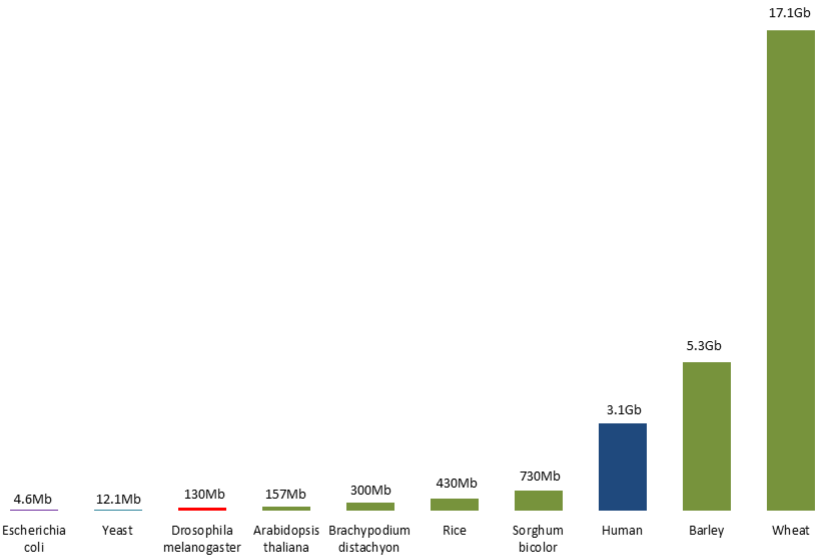

Figure 1.1: Genome sizes of selected plant and non-plant organisms. Mb = Megabase-pairs; Gb = Gigabase-pairs. Plant species are given in green color.

At the time of publication in 2000/2001 [21] the human genome sequence was reported to be the largest finished genome sequence with ~3,000 Mbp, achieved by a concerted financial and academic effort involving many different groups and institutions worldwide.

Many plant crop species equal or even largely exceed the size of the human genome, such as maize, barley and bread wheat, and remained unsequenced for a long time.

In the past, sequencing of (larger) genomes was a time-consuming and expensive task. With the introduction of next-generation sequencing technologies such as Illumina [22, 23] and Roche 454 [24], shotgun sequencing became a cost-efficient and fast alternative to traditional BAC-by-BAC sequencing approaches [25]. These NGS technologies typically generate short sequence reads of about 50-700 base pairs (depending on technology) from the genome sequence, often in very high coverage (meaning a specific position on the genome is covered by multiple distinct short reads) [26]. To reach longer sequence assemblies and, ideally, continuous pseudo-chromosome sequences, overlapping short reads are assembled by dedicated algorithms such as Velvet [27], Abyss [28], Newbler [29], ALLPATHS [30] and many more [31].

1.2.2 Plant genomes are formed by repetitive elements and whole genome duplications

A major factor which contributes to the formation of large genomes are repetitive elements ("repeats"). Transposable elements account for the predominating class of elements herein [32, 33].

LTR (*Long Terminal Repeat*) retrotransposons can be transcribed by reverse transcriptase and inserted back into the genome at a different place. Consequently, an enhanced activity of LTR retrotransposons can lead to a pronounced expansion of the genome size [34].

Repetitive elements can occur in thousands of copies in larger plant genomes and their multitudinous presence and high sequence identity can prevent assembly algorithms from joining adjacent sequences and introduce gaps in the genome sequence assembly instead [35]. Thus it is not only the genome size that makes larger genomes hard to sequence, assemble and analyse.

Whole genome duplications also contribute to the formation of large plant genomes [36, 37]. In fact, most modern plant genomes have undergone whole genome duplications (WGD) during their evolution as well as a number of additional genome modifications such as chromosomal rearrangements, fusions or loss of particular regions [38, 39]. For instance, there is evidence that a whole genome duplication took place in the genome of the common ancestor of the grass sub-families *Panicoideae*, *Pooideae* and *Ehrhartoideae* [40].

Gene sets that were duplicated by such an event can undergo different

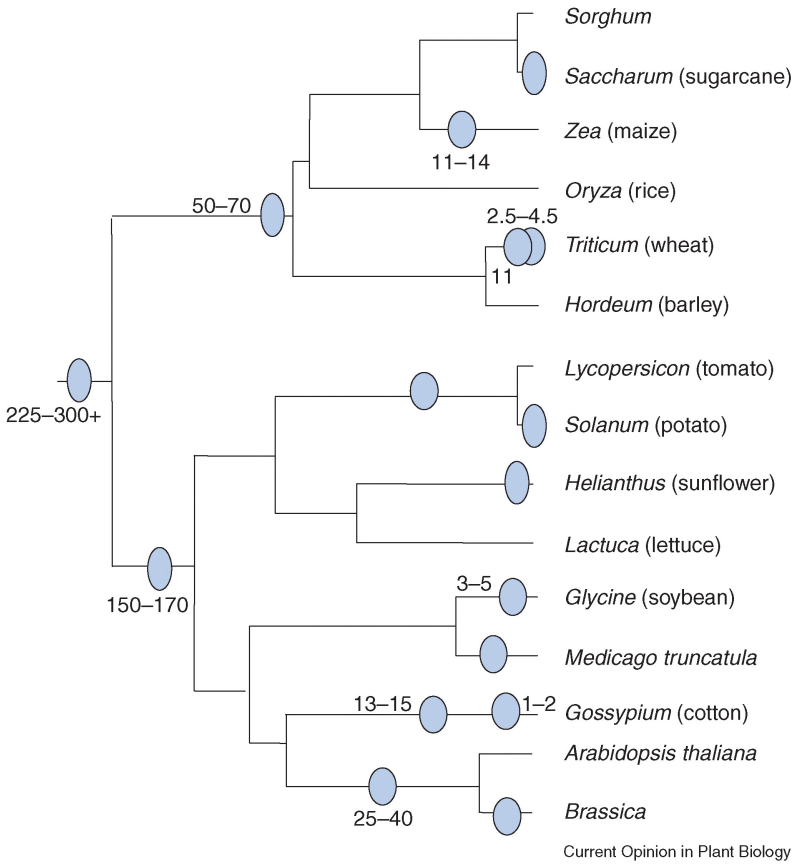

Figure 1.2: Polyploidisation events during the evolution of angiosperm plants. "Blue shaded ovals indicate suspected large-scale duplication events. Numbers indicate roughly estimated dates (in millions of years) since the duplication event" [37]. Figure and figure legend from [37], modified from [41], with kind permission from Elsevier.

evolutionary fates [42]. Due to the redundancy introduced by the WGD, duplicated genes can evolve towards new functions (sub-functionalization [43]) or degrade (pseudogenisation) without sacrificing the original gene function. Another possibility is that both copies of a gene are retained leading to an increased gene dosage.

Whole genome duplications and the resulting amplified gene set have a number of consequences and effects for an organism [44, 45]:

- with an additional gene set not under purifying selection, organisms may adopt to new environmental conditions and lifestyles by allowing random mutations in one of the copies without compromising presence or biochemical functionality in the remaining copy;

- the duplication (or multiplication) of a set of chromosomes and genes can promote the speciation of organisms as interbreeding with relatives or progenitors with deviating chromosome numbers may be handicapped or inhibited [46, 47];

- degraded/degrading genes (pseudogenes) and its domains can still provide the basis for genome innovation and the evolution of new genes, e.g. by bringing gene fragments into new genomic and regulatory context, mediated through retro-transposons;

Duplicated genes, however, can not only influence evolutionary processes on the genomic level but also on the level of transcription. While maintained on the genome sequence, duplicated gene copies may either be transcribed at the same level, leading to enhanced overall gene expression, or one or both of the copies may be transcriptionally depleted or silenced. Therefore, dosage effects associated with differentially transcribed gene copies may attribute to specific phenotypes and to speciation [48] and the adaption to certain environments and/or conditions as a consequence [49-51].

Whole-genome duplications as well as segmental duplications have been identified primarily from genomic regions showing significant homology between each other and duplication events could be dated using nucleotide substitution rates in protein-coding sequences [52].

Another important characteristic of plant genomes, polyploidy, is tightly associated with whole genome duplication events [37]. Whereas many of the sequenced reference plants with smaller genomes are diploid, many larger plant genomes are tetraploid, hexaploid or higher polyploid. However, even smaller genomes such as from *Arabidopsis thaliana* have experienced duplications during its evolution and remnants of polyploidy can still be identified [53, 54]. Among species with polyploid genomes, economically important crops such as potato (tetraploid) [55], cotton (tetraploid) [378] and bread wheat (hexaploid) can be found. Multiple sets of homeologous but not completely identical genes and non-genic sequences complicate genome sequence assembly and analysis. The genome of bread wheat consists of three different subgenomes (allohexaploid) with homeologeous genes showing a high average sequence identity around 97% [33, 379]. With many sequence assembly algorithms, this leads to the collapsing of most homeologeous gene sequences into chimeric contigs [291, 1, 380]. However, assembly and correct separation of homeologeous genes is critical for the development of specific markers and in breeding applications as it has been shown that different homeolo-

geous genes may contribute differently to important agronomic traits [90, 381]. One step further, if separate homeologeous gene assemblies could be generated, these cannot be directly attributed to their subgenome origin nor allocated to particular chromosomes. This would require the isolation, tagging and separate sequencing of subgenome chromosomes (as done by the IWGSC, see sections 1.3.1 and 4.6 for details) or novel strategies such as the comparative genomics approach described in this study [1].

1.2.3 Model plant genomes

As a consequence, until recently sequencing of plant genomes focused on crops and model plants with diploid and smaller to medium-sized genomes. Model (or "reference") plants are species "representative" for specific plant tribes and often show characteristics beneficial for work in experimental laboratories (such as short generation times, transformability etc.). Some model plants were selected for its close relationship to crops which have a larger and/or more complex genome [17]. Examples for model genomes are:

Arabidopsis thaliana, with its genome fully sequenced as the first plant in 2000 [13], is still the most important model plant system, e.g. for studying plant development, biological and molecular pathways and plant phenotypes. Its relatively small genome of ˜125 Mbp also supports both large-scale and in depth *in-silico* analyses and consequently can be considered the "best" analysed and described plant genome to date.

Arabidopsis thaliana is a member of the clade of the *Brassicaceae*, a family within the *dicotyledonous* plants. The group of dicotyledonous plants includes crops such as tomato, potato, soybean as well as all tree plants, whereas all grass species belong to the group of *monocotyledonous* plants. The first genome completely sequenced from the monocotyledonous group was rice (*Oryza sativa*) in 2005 [14], both a highly important crop and a model plant system.

For the monocotyledonous family of the *Poaceae*, where all economically important Triticeae crops such as wheat and barley belong to, *Brachypodium distachyon* was established as a model system due to its moderate genome size of 272 Mbp and diploid genome structure. In 2010, the finished genome sequence of *Brachypodium distachyon* was published [17], shedding new light on the evolution of grasses and enabling comparative genomics studies between *Poaceae* and non-*Poaceae* species. The *Brachypodium* genome is considered as a blueprint for the larger and more complex cereal genomes and

serves an experimental model system as well as a genome model.

1.2.4 Plant genome characteristics – conserved gene order

An important characteristic of grasses and *monocotyledonous* plants in general is the finding of long stretches of conserved gene order when comparing the genome sequences of related species [40, 56]. This feature, called synteny, makes comparative studies with less complex but closely related model organisms a valuable tool [57]; it has been shown that information about a gene in a model organism (such as localization) can be transferred to the crop if the homologous/orthologous genes are within syntenic regions [58-62]. This strategy is particularly promising for the identification of gene locations for traits of interest in complex grass genomes like those of wheat and barley.

The GenomeZipper concept makes use of the extensive syntenic relationships between the grass model organisms *Brachypodium*, Sorghum, rice and the complex cereal genomes barley, rye and wheat to construct virtually ordered gene maps for these crops [63, 64].

Syntenic relationships between genomes can be identified by various approaches. Historically, molecular markers (such as RFLP marker) and anchored ESTs gave evidence for strong syntenic relations within and between the grasses [65-70]. However, nowadays finished genome sequences are the easiest way to identify conserved gene orders.

Nevertheless, even in overly well-conserved syntenic regions and/or genomes, gene insertions, deletions, duplications and translocations can introduce local changes in the sequential order of genes [69, 71-73]. Model systems therefore cannot fully represent the actual gene content nor the accurate position and ordering of genes along chromosomes in crop plant genomes.

Finished whole genome sequences containing annotated genes overcome these limitations. They provide an overview over the almost complete gene repertoire of an organism. With a full genome sequence in hand, candidate genes underlying a particular trait or involved in a pathway/function can be identified even if they are not located in syntenically conserved region; moreover, molecular markers can be directly derived at low cost from the genome sequence resulting in a dramatically increased marker density.

In the absence of finished whole genome sequences especially from the highly complex cereal genomes of barley and wheat, model systems as well as

synteny-enabled approaches such as the GenomeZipper can act as extremely useful intermediate information resources on the way to fully sequenced crop genomes.

1.3 Triticeae and grass genomes – challenges and evolution

The genomes of many important cereals including bread wheat and barley bear great challenges for sequencing and analysis due to their large size, high repeat content and complex genetics.

With 5.1 Giga-basepairs (Gbp) in size, the genome of barley is almost double as large as the human genome (˜3 Gbp). The barley genome is diploid (2n) with a total of 7 chromosomes for which long and short arm are usually distinguished.

A repeat content of 84% is estimated for the barley genome; the overall high repeat activity and whole genome duplications in Triticeae ancestors are considered as major factors that contributed to the large genome sizes of many modern cereals in general [2].

It is thought that the common ancestor of both wheat and barley - as for all other cereals - contained five chromosomes, followed by a whole-genome duplication about 50-70 MYA and further evolving towards an intermediate ancestor with 12 chromosomes [40]. From there, the genomes of modern Triticeae were shaped by fusions of chromosomes or chromosomal segments [40], finally resulting in 7 chromosomes found e.g in barley, wheat and rye [74].

Archeological evidence indicates that both barley and wheat were cultivated by man since 10,000-13,000 years, being a very important factor for the establishment of permanent human settlements [75-78]. Cultivation, breeding and selection directly impacted the genomes of crops. In addition to selective pressures, hybridization of different species may introduce changes to the number of chromosome sets within an organism. These changes may lead to different levels of polyploidy, also resulting in an overall increased genome size.

As an example, the hybridization of diploid goat grass (*Aegilops tauschii*) with tetraploid emmer wheat (*Triticum dicoccoides*) gave rise to modern hexaploid bread wheat [79].

With a total size of ˜17 Gbp the genome of bread wheat is among the largest genomes sequenced and analysed so far. A repeat content of ˜80% is

estimated for the wheat genome, with primarily retroelements contributing to this [80].

The genome of bread wheat is composed of three different diploid subgenomes and is thus allohexaploid (6n) [81]. The subgenomes of modern bread wheat were contributed by three different grass progenitor genomes. Extant relatives of these progenitor genomes have been identified as:

- *Triticum urartu* as a close relative of the progenitor for the A subgenome [81-83]
- An unknown species likely from the *Sitopsis* section (which includes the species *Aegilops speltoides* and *Aegilops sharonensis*) for the B subgenome [84-86]
- *Aegilops tauschii* as the likely progenitor of the D subgenome [81, 87]

Hexaploid bread wheat originated from hybridization of cultivated emmer wheat (*Triticum dicoccoides*; tetraploid with A- and B-subgenome) with goat wheat (*Aegilops tauschii*; diploid with D-subgenome) in the Middle East about 8,000-10,000 years ago [76, 88]. The first appearances of tetraploid wheat strains (*T. turgidum*; A- and B-subgenome) were dated back to less than 0.5 million years ago [77].

Figure 1.3 provides a schematic overview about the genome evolution of modern bread wheat.

Comparing two different groups of bread wheat – wild and domesticated groups – identified significantly reduced nucleotide diversity in domesticated forms compared to ancestral lines. As a consequence, major domestication bottlenecks were hypothesized for the evolution of bread wheat and, even more severe, for the evolution of durum wheat (A- and B-subgenome containing) [78].

However, due to the lack of a wheat reference sequence and analysis concepts, nucleotide diversity and the frequency of single nucleotide polymorphisms (SNPs) between the subgenomes of bread wheat and its homeologous genes have not been investigated on a genome-wide level until recently [1, 90]. An average sequence identity around 97% was reported in previous studies for the homeologous genes in bread wheat, with some variation for different classes of genes [379].

With its hexaploid genome architecture, the bread wheat genome in principle contains three gene copies for every individual homeologous loci. However, homeologous genes may be subject to various fates including pseudo-

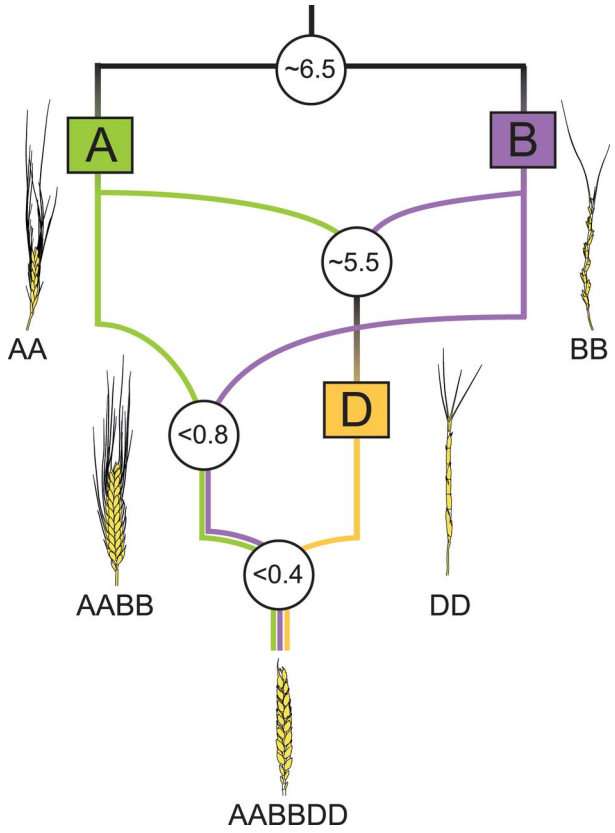

Figure 1.3: Model of the phylogenetic history of bread wheat (*Triticum aestivum*; AABBDD). "Approximate dates for divergence and the three hybridization events are given in white circles in units of million years ago" [89]. Figure and figure legend from [89], with kind permission from the American Association for the advancement of science.

genisation, neo-functionalisation and duplication, among others. Up to now, no genome-wide estimations on gene retention rates of homeologous genes in bread wheat were available. As described earlier, high repeat contents are a major problem for the assembly of genome sequences from short reads into longer scaffolds or even pseudo-molecules, due to the collapsing of highly similar or identical sequences into chimeric contigs. Polyploid genomes even increase this difficulty by duplicating or triplicating the amount of similar or identical sequences in the genome. A number of studies recently adressed the issue of assembling and separating homeologous genes in polyploid wheats, mostly using transcriptome data [291, 90]. However, apart from laborious and costly chromosome sorting strategies (e.g. using flow cytometry, see sections 1.3.1 and 4.6 for details), no methods for the genome-wide assembly,

separation and classification of homeologous genes in polyploid wheats have been proposed so far. In order to answer open questions like gene retention and nucleotide diversity in polyploid wheat and construct gene families, one of the major objectives of this thesis is the identification and elaboration of concepts suitable for the genome-wide assembly, separation and classification of homeologous genes in polyploid wheats using high-throughput next generation sequencing data.

While individual gene families such as genes involved in host-pathogen interactions [91, 92] were analysed before no systematic and comprehensive (multi-) gene family analysis on a genome-wide level has been conducted for both wheat and barley. Using the genome sequence resources generated in the sequencing consortia introduced in the next chapter, gene families will be constructed and analysed in the frame of this study for both the barley and the wheat genome with respect to and in comparison with genes from closely related reference organisms such as *Brachypodium* and rice. This analysis has been shown to help understanding the specific biology of an organism or a tribe by identifying expanded or contracted gene families and/or species- and/or lineage-specific genes. Chapter 1.5 provides more details and references for this as well as an introduction into the objectives, concepts and methodology of computational gene family analysis.

1.3.1 Triticeae genome sequencing initiatives

As genome sequences and embedded genes are valuable information resources for e.g. research, breeding and map-based gene isolation, genome sequencing initiatives for wheat and barley were initiated some years ago. The genome sequence resources generated within the international consortia introduced here are the basis for the analyses of the genomic repertoires in Triticeae carried out in this thesis.

The International Barley Sequencing Consortium (IBSC) [93] and the International Wheat Genome Sequencing Consortium (IWGSC) [94, 95] were initiated in 2006 and 2005 with the intention to coordinate and stimulate projects, efforts and funding, leading towards (near-) finished reference genome sequences for these two important crops for the scientific communities and for applied research. With the sequencing technologies available at that time, the timeframe for sequencing the genomes of barley and wheat was estimated to be several years, involving significant costs and manpower especially for the finishing of chromosome sequences.

The initial sequencing strategy focused on the construction of comprehensive BAC clone libraries with consecutive sequencing of the Minimum Tiling Path (MTP) [93]. With rapid advances in sequencing technology (next-generation sequencing) over the last couple of years, however, the generation of whole genome survey sequences with high genome coverage became economically feasible [96].

Typically, state-of-the-art sequencing technologies such as Illumina [22, 23] or Roche 454 [24] platforms generate reads of ˜50-700 bp size which need to be assembled into longer contigs and scaffolds afterwards [97].

In the presence of a high proportion of repeated sequence as found in the barley and wheat genomes, these assemblies remain fragmented with low N50 values [98] and no association to, or position on chromosomes [99]. Genetic maps based on a genotype-by-sequencing approach exist for both barley and wheat [100]. Genetic maps with a high marker density can help to position and order contigs on longer scaffolds or pseudo-chromosomes but their generation is laborious.

To circumvent these problems that exist in cereal genomes, new strategies had to be developed to identify genes, their chromosomal position and to characterize gene families.

In this thesis, concepts are described for the analysis of the gene repertoire and gene families in Triticeae plants containing particularely large and complex genomes. The results of comparative gene family studies with related crops and model plants give new insights into unique characteristics of cereals and their genome biology and provide a fundamental new resource that will stimulate numerous further studies.

1.4 Taxonomy and economic importance of cereals[1]

Cereals are an integral part of our daily life - in the form of bread, bio-fuel or animal feed to name only a few - and have influenced human culture and lifestyle since more than 10,000 years [75-78]. All economically important cereals such as wheat, barley, millet, sweet sorghum, maize and rice belong to the family of *Poaceae* (sweet grasses), a diverse and large subfamily of the monocotyledonous flowering plants [102, 103]. In contrast to the dicotyledonous plants, to which e.g. *Arabidopsis thaliana* belongs to,

[1] section adapted and modified from Spannagl, M., master thesis 2009 [101]

monocotyledonous plants do not show any secondary growth in girth and their number of cotyledons is limited to one.

Sweet grasses are among the largest plant families with more than 10,000 species and 650 genera and they can be found in all climate zones around the world [103].

Within the *Poaceae*, three different sub-families can be distinguished which contain the most important cereals for human nutrition: *Panicoideae*, *Pooideae* and *Ehrhartoideae.*

Based on fossil evidences [104] and the comparison of plastid and ribosomal DNA between grass species [105, 106] it is thought that these three sub-families evolved from a common ancestor about 50-70 million years ago [103, 107].

The *Panicoideae* subfamily comprises the species maize, sorghum, millet and sugar cane whereas the different varieties of rice belong to the *Ehrhartoideae* subfamily. The *Pooideae* family can further be subdivided into *Aveneae*, *Poeae*, *Bromeae* and Triticeae which include the economically important cool season grasses. Barley, wheat and rye are the most prominent members of the Triticeae tribe [103, 107].

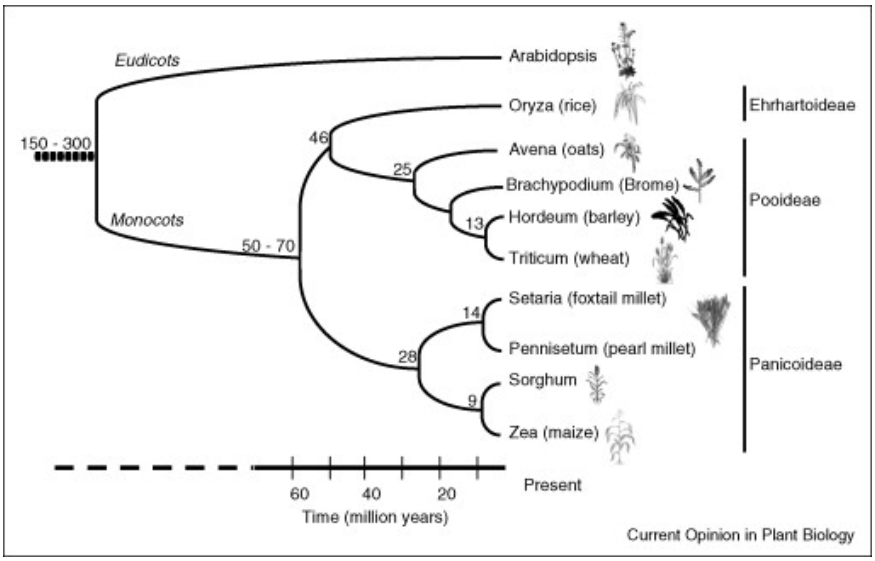

Figure 1.4: Schematic illustration of the phylogenetic relationships between cereals. "Divergence times from a common ancestor are indicated on the branches of the phylogenetic tree (in millions years)" [40]. Figure and figure legend from [40], with kind permission from Elsevier.

Grasses are of utmost importance for world human nutrition, both in form of its grains or as animal feed. Further applications include its use

as starch-, sugar-, oil-, and cellulose-resource and cereals such as sugarcane or bamboo gain more and more importance as renewable bio-ethanol and bio-fuel resources. Although the *Poaceae* are comprised of so many different species only a few are of greater economic importance. Many of the cereals harvested today are actually the results of multiple rounds of breed selection and crossing over thousands of years [75, 108-110]. During the "green revolution" more than 50 years ago, food crop productivity could be increased significantly, attributed especially to the development of cereals with a much higher grain yield [111].

Today, maize (*Zea mays*), wheat (*Triticum* varieties) and rice account for the top-3 of the most harvested grass crops world-wide [112] (not considering sugar cane with the highest overall production). Figure 1.5 shows the respective yields harvested in 2012 as determined by FAOSTAT [113].

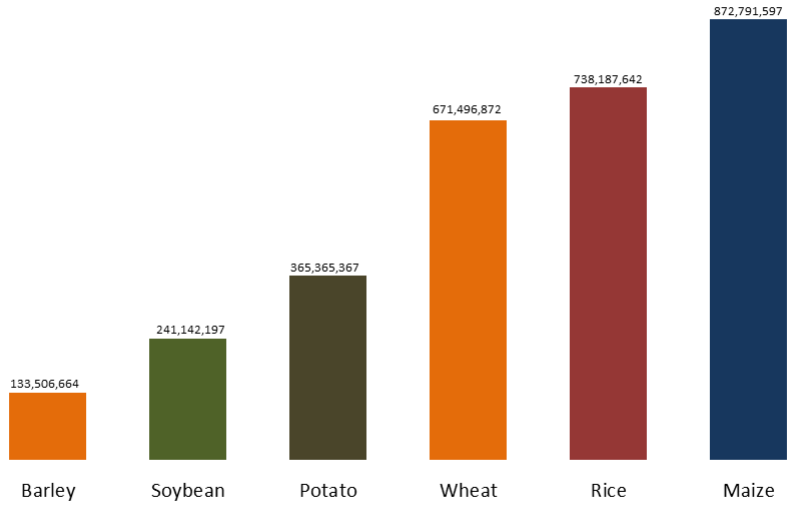

Figure 1.5: Food and agricultural commodities production as determined by FAOSTAT for the year 2012 [113]. This ranking includes selected crop plants only. Numbers given are in tons produced in 2012.

With a global harvest of ˜670 million tons in 2012 (FAO [112]), wheat substantially contributes to human nutrition, accounting for ˜20% of the calories consumed [112]. Wheat is grown as different cultivars around the

world, including bread wheat and durum ("pasta") wheat to name only a few.

In 2012, ~133 million tons of barley were produced (FAO [112]). Barley is primarily used as malting barley during beer brewing but is also of great importance as an animal fodder resource due to its relatively high protein content [114].

Both barley and wheat are grown in many different environments across the world. Barley is considered more stress tolerant than wheat [115] making it an important food resource for poorer countries where agricultural conditions often remain difficult and environments harsh [2, 116].

A number of great challenges have to be dealt with when cultivating croplands in the future. These include an ever-growing world population, climate change with desertification and other effects as well as the on-going industrialisation of emerging nations coupled with growing land consumption. The targeted breeding of important crops to change and adopt them to specific conditions and locations (such as dry habitats) plays a key role herein.

1.5 Concepts and methods for the analysis of genes and gene families in plants[2]

Within this thesis, gene families have been analysed for both the barley and the wheat genome with respect to and in comparison with genes from closely related reference organisms, namely *Brachypodium*, sorghum and rice. This analysis has been shown to help understanding the specific biology of an organism or a tribe by identifying expanded or contracted gene families and/or species- and/or lineage-specific genes. The following chapter provides an introduction into the objectives, concepts and methodology for the identification of conserved and species-specific gene models and the computation of gene families in plant genomes. Moreover, references and examples for gene family studies/analyses in other plant genomes are given and important findings are highlighted.

Whole genome duplications and other modifications, described in more detail before, may influence and change the gene content of an organism.

[2] section adapted and modified from Spannagl, M., master thesis 2009 [101]

All these changes and events may result in expansions of gene families but also in gene loss and in the birth of new genes through sub-functionalisation and gene fusions [117, 118].

However, it is not only the genome-wide mechanisms such as WGD that play a vital role in gene and gene family expansions and the formation of species-/lineage-specific genes and gene families but also (local) gene duplications, TE-mediated gene shifting [119] and horizontal gene transfers [120, 121]. Pseudogenisation describes the loss of function and gradual degradation of a gene model and accounts for the development of many species- and lineage-specific genes we observe today [122]. This is often put into effect by a gene accumulating random mutations which may disturb the open reading frame at some point or by the insertion of transposable elements into its sequence. Pseudogenisation events can be observed at a higher frequency when genes exist in higher copy number, e.g. mediated through gene and whole genome duplications, and at a greater level of functional redundancy as a result [37, 122, 123].

The identification of genes conserved between related species has been one of the main objectives in comparative genomics since decades but also species- and/or lineage-specific genes and gene families are of great interest for researchers. These genes and gene families contribute to the speciation of organisms and play an important role in the adaption to specific environmental conditions and defense mechanisms against pathogens [124].

On the other hand, many studies comparing genomes of closely related organisms report high numbers of gene pairs with overall conserved coding sequence, even if their genome sizes differ significantly [125]. The sequences of DNA histone proteins, for example, were shown to be well conserved even over different biological kingdoms [126].

If sequences of genes in related species appear to be conserved over a long period of time it is thought that they are under preserving selection pressure [127]. Homologous genes, sharing high sequence similarities between related species, are termed orthologous genes if they share a common ancestor and likely perform the same biological function in their organisms [128]. In contrast, fast evolving genes and gene families often appear related to resistance traits involved in defense mechanisms against plant pathogens such as fungi and bacteria [129-131]. Here, the capacity for genetic innovation is crucial for a plant to act against new evolving pathogens.

Genes accounting for specific traits of modern cultivated crop plants are of special interest in all agricultural applications. Such traits of interest

include the ability of specific ecotypes to adapt to dry habitats as well as tolerance against salty ground or the greater/lower harvest of a specific cultivar. Additionally, the identification of genes involved in pathways such as specific photosynthesis reactions (C3, C4) is another important task [4, 132].

The genes accounting for desired qualities such as drought tolerance or increased yield can, at least partly, be assumed in the portion of species- and/or lineage-specific genes of the respective organisms [133, 134]. Therefore, the identification and functional description of shared and specific genes and gene families is of great relevance. To modify specific traits such as the oil content in a plant for agricultural use, e.g. by targeted breeding, the genes involved in this characteristic are an excellent starting point. However, not only the presence or absence of genes or the genetic variation within may determine the formation of a specific plant trait but also several additional mechanisms potentially contribute such as transcription regulation, small RNAs, DNA methylation or histon modifications. Copy number in corresponding, orthologous gene families appears to be dynamic even between closely related species [135, 136]. Expansions or contractions in gene family size were identified in numerous genome comparisons and attributed to natural selection, resulting in new findings and hypotheses about evolution and functional repertoire of specific organisms or lineages [137-140].

Within this study, Triticeae- and species- specific genes and gene families (as well as expansions and contractions herein) are identified in the genomes of barley and bread wheat and analyzed for their potential functional role. To analyse for shared and specific genes and gene families between related organisms several methods and strategies have been proposed before. These were developed for and applied to a number of organisms and gene families, not only plants.

One of the first comparative analysis of gene families based on a complete genome sequence was published by Sonnhammer in 1997 [141]. In this analysis, gene models predicted on the finished genome sequence of *C. elegans* were compared for sequence similarity with previously known genes in human and *Haemophilus influenceae*. Additionally, nematode-specific gene families were identified by grouping genes according to their PFAM domains [142] into clusters. By analysing clusters with genes lacking any significant sequence similarity with non-nematode proteins in more detail, it was possible to assign putative functional descriptions to some of them.

Based on the identification of orthologous gene groups in the genomes of

prokaryotic organisms [135, 143, 144], the database Clusters of Orthologous Groups (COG) was established as a resource for orthologous proteins found between multiple species [145, 146]. COG cluster are computed using pairwise BLAST [147] searches between the protein sequences of fully sequenced organisms. Hereby, an orthologous pair is established if two protein sequences from different genomes show bi-directional best BLAST hits. If orthologous pairs are found between at least three different lineages a COG is annotated.

When computing clusters of orthologous groups (COGs) for the genomes of more complex eukaryotic organisms, such as yeast (*Saccharomyces cerevisiae*), three different observations were made:

- Generally, eukaryotic genomes exhibit significant more gene duplications which can cause wrong associations of best BLAST hits;
- Eukaryotic proteins are often composed of more than one functional domain and these can be arranged in complex order [148]. There are severe difficulties involved with sequence based search methods for detecting homologs of multidomain proteins [382]. This can be caused by a number of promiscuous, unspecific domains occuring together with more specific domains which can cause wrong associations in sequence homology searches between the domain architectures of proteins. Wrong links between otherwise unrelated proteins can also be established by domain-only matches, when sequence pairs share similarity due to the insertion of the same domain into both sequences [383].
- The genome sequences along with the gene predictions remain unfinished and incomplete for many eukaryotic genome sequencing projects. While this is the case, true orthologs are potentially missed in one or the other organism. Instead, incorrect ortholog associations may be made with sequences sharing second-best sequence homology (remote homologs).

To overcome some of these difficulties, in particular to be able to deal with frequent gene duplications also present in many plant genomes, alternative approaches have been developed which are capable to decide between so-called “young” and “old” paralogous sequences. Genes which were duplicated within an organism after the split of all species analyzed are termed “young” paralogs. These genes are thought to carry out the same or similar

biochemical functions within that organism. "Old" paralogous genes, on the other hand, are genes duplicated before the first split of the species analyzed and which putatively diverged into different biological functions afterwards [149]. Moreover, because of the eukaryots' complex domain structures, all methods had to be able to incorporate the global relationships of two protein sequences.

Both multiple alignments and phylogenetic trees can in principle be used to construct orthologous groups and discriminate between young and old paralogs. However, their computation is time- and resource- intensive, especially for larger datasets. As a consequence, more efficient algorithms had to be developed to compute groups of orthologous and paralogous genes for large datasets, often incorporating thousands of proteins from multiple species and lineages. These algorithms include INPARANOID [150], EGO [151] and OrthoMCL [149] as the most well-known representatives.

INPARANOID [150] utilizes BLAST to identify homologous protein sequences followed by the extraction of bi-directional best BLAST hits between two sequences to establish an orthologous group. Subsequently, multiple rules are applied to identify paralogs originating from gene duplications after the split of two species (termed "in-paralogs" here). This method has been successfully applied to protein sets from yeast and mammals where a good accordance of orthologous groups computed with INPARANOID with manually curated gene families could be observed. However, as a consequence of its rule-based methodology, INPARANOID can only be applied to two distinct protein datasets at the same time. This is a severe limitation of the concept, especially when protein data sets from multiple species or lineages need be analysed in one study. To overcome these limitations, MultiParanoid [152] was developed as an extension of INPARANOID. Here, the multiple pairwise orthologous groups computed with INPARANOID are being merged into orthologous groups of multiple species using a clustering algorithm. Only groups of orthologous genes are merged which share the same common ancestor.

EGO [151] is a method to compute orthologous gene groups on TIGR gene indices [153, 154] using a similar approach as the Computation of Orthologous Groups – COG. EGO can be readily applied to the gene datasets of multiple species, but it inherits the same limitations as already discussed for COG.

OrthoMCL [149] is a widely used method to identify groups of orthologous genes in the genomes of eukaryotic organisms. While the strategy is

similar to that of INPARANOID, protein datasets from multiple species can be analysed directly with OrthoMCL. To distinguish young paralogous genes from older gene duplications that occured before a species split, OrthoMCL utilizes the following concept: "Young" paralogous sequences are being identified and grouped together with orthologous genes whenever there is another gene with greater sequence similarity in the same organism than it is in all other species compared. Sequence similarities are computed using BLAST and relationships between sequences are established in a bi-directional way. After that, a graph is constructed where proteins are represented as nodes and the weighted edges correspond to the sequence similarities between the proteins. This graph is then being clustered with the Markov Clustering Algorithm MCL [155]. MCL computes random walks through the graph determining regions of high flux and connection (the clusters) which can be separated from regions with low or no connections. OrthoMCL (and its variant MCLBLASTLINE) has been used in a number of genome analyses to determine gene families shared by multiple species, e.g. in the comparative analysis of the genome of *Phaeodactylus* (duckbill platypus) [156], for the plant genomes of Sorghum [4], tomato [15], *Brassica rapa* [157] and cotton [6] as well as for the fungal genomes of *Sclerotinia* and *Botrytis* [158]. OrthoMCL is one of the major tools used in the gene family analyses of cereal genomes outlined and discussed in this thesis.

1.6 Genome databases and plant genome resources: an overview

Within this thesis, novel methods were developed and applied to the genome sequence data from polyploid wheat to assemble, separate and classify homeologous genes. Gene families have been constructed and analysed for both the barley and the wheat genome with respect to and in comparison with genes from closely related reference organisms such as *Brachypodium*, sorghum and rice. As a result, heterogenous, high-volume and complex data had to be integrated from different resources and managed in a dedicated database framework as well as disseminated to the public through specialized tools and interfaces. This step is of great importance not only as a prerequisite for efficient genome data analysis (as performed in this study when constructing gene families, managing versions and integrating heterogenous data) but also for the usability of the newly created Triticeae

genome resources by experimental biologists and breeders. As an example, the representation of the wheat gene sub-assemblies together with their reference genome association and subgenome origin (see chapter 3.2 for details) asks for both entirely new web and search interfaces and internal storage. This chapter aims to provide an overview of existing genome database systems and outlines the specific needs for the integration, management and dissemination of the data types generated (not only) in this study. This chapter also introduces the PGSB PlantsDB database system which was enhanced and used for the integration, management and dissemination of the Triticeae genome data described before.

The plant genome sequencing projects introduced before as well as multiple studies building on top generate massive amounts of both raw data and project results. It is crucial not only for the plant research communities to store/archive, manage, integrate and visualize these data. Hereby, several main objectives for the management of plant genome data can be identified:

a.) Archiving and versioning of raw genomic data such as WGS short read sequences and single nucleotide polymorphism (SNP) annotation.

b.) Storage and integration of project and analyses results such as gene predictions with whole-genome sequence assemblies, functional annotations, genetic and physical maps (markers) etc.

c.) Visualization of data via web-accessible platforms and provision of specialized tools to further analyse and mine data, often in the context of other integrated data.

Thanks to the cost-efficient next-generation sequencing technologies (described above) the amount of raw sequence data generated, not only in plants, has been growing significantly over the last few years [159-161]. In order to meet the objectives for data management, integration and visualization the associated storage capacity has to grow simultaneously. As an alternative, data compression algorithms and efficient data structures have been investigated especially for raw genome sequence reads and are in use at the major sequence archives Genbank and EBI [162, 163]. One step further, Cochrane et al. propose a graded system for submitting sequence data to the public archives considering ease of reproduction and sample availability when choosing a compression level [164].

Figure 1.6 illustrates the trend of sequence data stored at EMBL-Bank (operated by the European Bioinformatics Institute, EBI) over the last decades.

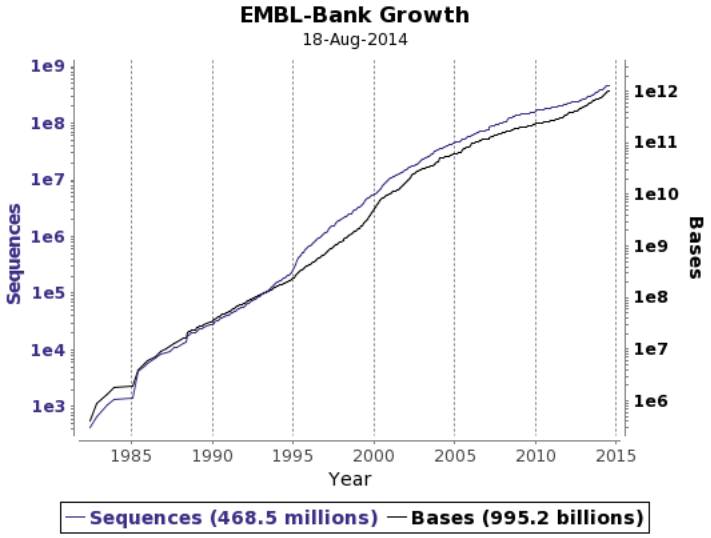

Figure 1.6: Data growth within the EMBL-Bank from ˜1980 to 2014. Figure from [165].

Not all tasks in the management of biological data are/can be usually addressed by a single center or institution, which is especially true for plant genome research. For data management and storage, genome data can be categorized in two different ways:

a.) by the type and nature of data, such as raw sequence reads, gene predictions, genetic maps etc.

b.) by its biological origin, namely the species.

As a consequence of the growing amount of genome data, the International Nucleotide Sequence Databases (INSD) [166] consisting of GenBank (hosted by NCBI, US, from 1982) [167, 168], the DNA Databank of Japan (hosted by DDBJ, Japan, from 1987) [169] and European Molecular Biological Laboratory (EMBL; hosted by EBI, Europe, now the European Nucleotide Archive - ENA, from 1982) [170, 171] were established to serve as central data archives for published or publicly available genome data across the biological kingdoms. These data archives were designed to accept submissions of raw and processed genome data from any institution through standardised web forms and protocols. Both ENA and Genbank provide a rich set of interfaces to search, query, browse and download data and both resources are set up to deal with multiple versions of a dataset, such as updated/improved genome sequence assemblies from the same species. EMBL and Genbank synchronize their data content daily to ensure maximum data

consistency but also to provide a certain level of redundancy in the case of technical failures. Both ENA and Genbank consist of multiple sub-units or databases which are focused on different types of data. Examples are the Short Read Archive, resp. Sequence Read Archive (SRA) [172] for the submission and archivation of raw sequence reads from NGS projects or EMBL-Bank [173] for the submission of genome annotation.

It has become common standard to submit all raw data from a genome sequencing project, including raw sequencing reads to the respective ENA or Genbank instance before or with the publication of the corresponding study. Given the rapidly growing amount of sequence data all archives have to deal with great challenges in data storage and analysis capacity [160]. As computer storage facilities cannot grow with the same pace as the sequence data (due to technical and cost reasons) at this time [164], data compression is a vital concept to cope with the data and new data compression algorithms are subject to further research [174].

Along with the raw data, analysis results such as genome assemblies, gene predictions etc. can be submitted to the central repositories as well. However, both EMBL and Genbank cannot provide all the views, tools, data and integration levels that user and research communities such as plant breeders require to assist their daily research and reach their goals. Among other features, ENA and Genbank e.g. do not provide genome browsers (such as Gbrowse [175]), community annotation interfaces or comparative genomics tools and views.

User communities often have very different requisites for online genome data resources, not only between biological kingdoms such as plants and bacteria but even differing from species to species. As an example, biologists working with monocotyledonous plants (to which the Triticeae belong) may be more interested in synteny visualization tools as the gene order is more conserved here compared to the dicotyledonous plants [176]. As a consequence, more specialized genome resources and databases were developed independently for many kingdoms, tribes and species. However, to avoid duplicated efforts in the development of tools and interfaces, many genome databases and resources rely on publicly available or shared components or tools such as common databases schemas (CHADO [177]), visualisation tools (GBrowse [175], JBrowse [178], Apollo [179]) or even complete software suites (GMOD [180]).

Some of the most recognized genome data resources and databases include:

- Flybase (http://flybase.org) [181, 182] was established as an online (genome) database and annotation system for the insect model organism *Drosophila melanogaster*, a species widely used for (developmental) genetics since decades. Since its initial setup in 1993, numerous tools were developed to assist the *Drosophila* research community and new data (types) such as variation calls and genome sequences from additional *Drosophideae* species were integrated. Flybase is actively involved in many ontology, bibliography and Model Organism Database (MOD) initiatives and receives on-going funding and support from the US and UK.

- Wormbase (http://wormbase.org) [183, 184] was founded in 2000 to provide a platform for the genetics and genomics of *C. elegans* and related nematodes. Numerous tools such as Gbrowse, WormMart (a BioMart data warehouse system) or Synteny browsers were integrated into the database framework as well as lots of metadata collected such as disease ontology terms, publications, motifs and laboratories involved in the project.

- The UCSC genome browser [185] hosts a collection of genomes including the latest genome sequence assemblies of human [186], chimp and mouse. Several tools such as Genome Browser and Gene Sorter provide access to data which includes expression data, gene predictions, in situ images and many more. The UCSC Genome Bioinformatics Site also provides portals for both the Neandertal project [187] and the ENCODE project (The Encyclopedia of DNA Elements) [188], an effort to „build a comprehensive parts list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active“ [189].

- The Arabidopsis Information Resource (TAIR) [190, 191] is a database resource dedicated towards the collection, integration and curation of genetic and molecular biology data from the model plant *Arabidopsis thaliana*. TAIR was established in 2001 after the full genome sequence of the first plant organism *Arabidopsis thaliana* was published in Nature [13]. The TAIR database was set up as a central data hub to collect, integrate and curate the growing amount of (heterogenous) *Arabidopsis* data generated in various labs and provides intuitive views and analysis tools for different data types. These include

„complete genome sequences along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, mutant informations, publications, and information about the Arabidopsis research community“ [192]. Lately, TAIR also integrated the genome sequences of additional *Arabidopsis* varieties and provided access to the resulting SNP callings [12]. TAIR provides an exceptionally high degree of active data curation by experts (such as gene structures) and profits greatly from frequent community data submissions (such as gene product function updates). TAIR is being updated regularly and incremental data builds are released to the public about 1-2 times a year (termed 'TAIR8', 'TAIR9', 'TAIR10' and so on). For the initial release of TAIR, MIPS MAtDB (The MIPS *Arabidopsis thaliana* database, an instance of PGSB PlantsDB) [193] contributed all data collected so far herein and mirrors some of the TAIR data content in its interfaces from then on. TAIR is being operated in the US by a consortium of different institutions, however, institutional funding for TAIR recently stopped in 2013 but there are attempts to (at least) preserve the current data inventory and functionality by moving the different TAIR components to servers from the iPLANT initiative [194] and establishing a subscription model [192].

For biologists working with the model plant *Arabidopsis thaliana*, the TAIR database makes a good case for a "1-stop-shop". Almost all existing information about this particular organism can be found here at a single location with a high level of data integration, even stocks can be ordered online [192]. Similar, albeit less data-rich, genome resources are available e.g. for rice [195] and tomato [196]. However, TAIR and other databases focused on individual species usually do not collect and provide data about other species and their relationships between each other. Having data from multiple (related) species in one database system can help to address a number of important biological questions. This includes the identification and representation of species- or lineage-specific genes, conserved genes (homologous/orthologous genes), gene families or conserved gene orders along chromosomal stretches between species.

As a consequence, additional genome resources were set up which typically focus on specific biological kingdoms (such as plants, bacteria etc.) or tribes (such as legumes or cereals in plants) and aim to integrate genome data from multiple species. For the plant kingdom, some well-established

cross-lineage resources include:

- Ensembl Plants [197] is the plant division of Ensembl Genomes, a genome database framework developed by EMBL-EBI in UK. Funding for Ensembl Plants is provided from different institutions and grants in which other plant genomics and bioinformatics groups contribute to the development of Ensembl Plants. One of the key components of Ensembl Plants is the embedded genome browser allowing the integration and concerted visualisation of various genomic data types such as genes, alignments, synteny, resequencing data, markers and many more. At this time, Ensembl Plants hosts genome data from about 25 different plant species from almost all major tribes.

- Phytozome [198] is a plant genome resource operated at the Department of Energy's Joint Genome Institute (JGI) and the Center for Integrative Genomics in the US. Phytozome hosts genome data from many plant species where JGI was actively involved in the genome sequencing, annotation and data analysis. Special focus is given to the provision of tools and views to promote comparative genomics between included plant species, e.g. by constructing families of paralogous and orthologous genes. These analyses particularly benefit from the genome data from a large number of different/related species and a great coverage across the plant phylogenetic tree. As of release version v9.1, genome sequences and gene annotations for a total of 41 different plant species are available from Phytozome.

- Plaza [10] is a plant genome resource with a special focus on comparative genomics and evolutionary analyses. With the latest release version 2.5, Plaza integrates the genome sequences and structural and functional annotation of 25 different plant species. The data can be mined, analysed and visualized with various tools including plot visualization, Ks-graphs and gene family clustering. Plaza is run by the Bioinformatics & Systems Biology unit at VIB/Ghent University in Belgium.

- Additionally, there are many genome database resources focusing on genomic and molecular data from specific plant tribes or lineages. These databases are often tightly connected to their respective user communities and often serve both as data analysis and data management centers. Examples include the Legume Information System (LIS)

[199] for the *Fabaceae* plant family, Gramene [200, 201] for the grasses, GnpIS [202] for plant and fungi genomes analysed by the INRA centers in France, SOL Genomics Network [196] for the *Solanaceae* plant family.

PGSB (formerly MIPS) PlantsDB [203] also belongs to this class of plant genome resources although its key aspects and motivation differs from many databases introduced before. The mission of PGSB PlantsDB includes two main objectives:

- PGSB PlantsDB was established to serve as an informative, intuitive and data-rich interface to the research community for selected plant species and its genome data. The decision what species to include is driven by the importance of a species/data set for plant genome research as well as by active participation/involvement of the PGSB group in the analysis or/and data management of a particular plant species or family.

- PGSB PlantsDB and its individual components serve as a data management center for all plant genome projects finished or underway at the PGSB group. Analysis results as well as raw genomic data need to be archived, integrated and versioned in any mid- to large-scale scientific environment even if the results of a particular project or analysis may not be displayed to the outside world. PlantsDB tools and components also contribute to many plant genome analyses by providing custom datasets (such as upstream or promotor sequences), genomic views (synteny) and gene family results.

PGSB PlantsDB was initially developed as a plant genome database for the first model plant organism *Arabidopsis thaliana* in 2001 [193, 204]. With the fast growing number of finished plant genomes and new data types, PlantsDB was continuously extended with new data modules, interfaces/views and tools. For that, the design of PGSB PlantsDB with a modular architecture and a high degree of data normalization from the beginning showed to be very beneficial.

A special focus in further developments of PlantsDB was given to tools and interfaces for comparative genomic analyses in plants. Examples include the visualization of conserved gene order between genomes (synteny; CrowsNest tool [3]), computation of orthologous gene families across many

related plant species (via OrthoMCL [149]; results in gene reports) and integration of external resources such as SIMAP (Similarity Matrix of Proteins) [205, 206]. Another field of major interest is the integration and visualization of all data and analyses results gathered and produced within the Triticeae and grass genome projects introduced before. This includes the integration and visualization of the barley and wheat GenomeZipper data and results [64] as well as the barley physical and genetic maps [2]. A dedicated PlantsDB instance was also developed to serve all data from the UK wheat genome sequencing project [1].

At the date of writing (April 2014), PGSB PlantsDB hosts genome data from 11 different plant species in its public domain and many more in private or password-protected instances.

1.6.1 Towards the interoperability between (plant) genome databases: objectives and concepts

---

As a consequence of international sequencing consortia and shared efforts in deciphering the complex genomes of barley and wheat (see chapter 1.3.1), resulting genome sequence data and analysis outcome usually resides in distributed database resources at different institutions. When working with this data, data acquisition and integration, e.g. for downstream analyses, can be very challenging, up to the point where expert knowledge is required. Data aggregation from distrinuted resources played an important role both for some analyses of this study as well as for making the data generated here useable by the broader research community. As an example, wheat gene sub-assemblies (with their subgenome prediction) were provided by PGSB as search indexes to be integrated with SNP data from the UK wheat consortium in the Ensembl Plants database system [197]. This allowed for a straightforward inspection and evaluation of the subgenome assignments in the context of additional evidence originating from another resource [197] without the need of on-site data curation or integration. This chapter describes objectives and technologies to aggregate genome data from distributed genome resources and databases and introduces the projects and solutions implemented for the use of distributed Triticeae genome data in the framework of PGSB PlantsDB and this thesis.

As outlined before, plant genome resources and databases operate on

different core areas and with distinct objectives, both in terms of plant species/lineages incorporated and tools and views provided to end users. Although there is some redundancy (tools, plant species), the major plant genome resources are very much complementary in their offering of plant genome information to researchers. As a consequence, and as there are no plans for a single centralized plant genome resource, attempts were made to establish or improve interoperability between existing plant genome resources [207, 208]. The main advantages of cross-linked databases for end users and developers include:

- A seamless search and navigation experience for users if as much data as possible is provided for a query within a single framework. The complexity of the query and the fact that the data is actually distributed among different physical entities/partner databases is hidden from the user.

- Enables developers to include additional data and/or species into existing tools and views to enhance power (e.g. in phylogenetic analysis tools or synteny viewers) without the need to take care about data integration and curation at the local side.

- Helps to avoid duplicated efforts in data integration and curation for the same data at different genome resources.

A number of different projects was initiated over the last couple of years to address the logic and technical challenges of plant genome database interoperability. Some of the problems identified include different database schemas, choice of applicable communication technologies and missing/incomplete or incompatible ontologies and controlled vocabulary.

One of the first initiatives which aimed to develop and establish technology for genome database interoperability was the BioMOBY project [207, 209], starting in 2001 with members from many different countries and institutions. BioMOBY provides a registry for web services enabling the interoperability between resources for genomic data. BioMOBY services can be invoked directly within code and therefore also facilitate the interoperability between analytical services and tools and the setup of (remote) bioinformatic workflows, e.g. within workflow management tools such as Taverna [210]. Three ontologies are defined by the BioMoby project, describing biological data types, biological data formats and bioinformatic analysis types.

The BioMoby registry, where all services are registered with their meta-information, can be queried using BioMOBY clients such as Gbrowse Moby [211].

PGSB PlantsDB participates in the BioMOBY project and provides some of its core functionalities as BioMOBY web services. This includes services to retrieve gene and gene product annotations as well as gene and genome sequences.

In 2011, the European Commission funded a project within its 7th framework programme, transPLANT (trans-National Infrastructure for Plant Genomic Science), to define common data exchange formats, ontologies and standards and to establish a trans-national infrastructure for plant genomic science between major European plant genome resources. In the framework of transPLANT, a search interface (e.g. to search for a gene function) was set up at www.transplantdb.eu which integrates result hits from Ensembl Plants [197], PGSB PlantsDB [3], GnpIS INRA Versailles [202] and IPK Gatersleben [212], providing a first prototype of a virtual one-stop plant genome resource. This query interface is complemented by a searchable registry of plant genome resources which are collected worldwide.

A similar approach is taken by the Distributed Annotation System (DAS) [208, 213], an environment which allows multiple partners to contribute to and exchange gene predictions and annotation for a specific organism. With DAS, annotation information gathered from different places/institutions do no longer need to be integrated into a single, centralized database but can remain at the partners side, ensuring full (local) control over the data and eliminating data integration issues. DAS clients enable the aggregation of data distributed over multiple partners and provide the user with single integrated views, e.g. via a website or genome browsers such as GBrowse or Ensembl.

Another strategy to avoid redundant administration and data curation efforts is to implement and use components which can be/are shared by multiple partners. This concept is often used with new genome projects and has the advantage that all data produced by collaborating partners will be integrated into a single resource from the start, eliminating the need to use interoperability technology later on. One of the projects making use of this concept is „GMOD in the Cloud“ [214], a cloud implementation of the widely-used Generic Model Organism Database (GMOD) [180] software components. By running GMOD components such as the CHADO database [177], GBrowse [175], JBrowse [178] and Apollo [179] or Web Apollo [215]

in a cloud environment, multiple users can contribute data to and/or access pre-configured genome database tools and views without much effort in installation and administration. Therefore, „GMOD in the cloud“ could ideally be employed at annotation jamborees or by communities/labs without much bioinformatic infrastructure, but it can also serve as a stand-alone genome database framework for non-distributed work e.g. if flexible connectivity is one of the priorities.

Web Apollo [215], which is also part of the GMOD framework, is a browser-based tool for the visualisation and editing of sequence annotation. Unless its stationary edition Apollo [179], Web Apollo facilitates distributed community annotation, where multiple researches from different locations may be working on the same sequences at the same time. Web Apollo, as all other tools enabling distributed editing and database access, needs to keep users in sync (e.g. by real-time updates) to ensure data consistency and manage user grants and access privileges.

Although cloud genome database frameworks and tools are limited by a few factors such as cloud computing costs, scalability and missing direct control over the computing hardware, the concept appears to be a welcomed alternative especially for smaller-scale experimental labs and communities producing genome sequence data.

Chapter 2

Material and Methods

2.1 Comparative analysis of gene families in complex cereal genomes

To compute clusters, or groups of orthologous genes and to define gene families from there, the OrthoMCL software [149, 216] in version 1.4 and version 2.0 was used. OrthoMCL utilizes BLASTP [147] to construct a matrix containing pairwise sequence similarities between all input protein sequences. In all OrthoMCL analyses described in this thesis, an e-value cut-off of 10e-05 was used for the BLAST search. To obtain the final cluster structure and to define orthologous and paralogous relations, the Markov clustering algorithm [155] is applied to the graph connected by sequence similarities between proteins. An inflation value (-I) of 1.5 was used with all OrthoMCL analyses described in this thesis which corresponds to the OrthoMCL default for this parameter, shown to perform best in studies with eukoryotic data sets [149].

In order to avoid tight clustering of very similar or identical protein sequences originating from the same genomic locus, splice variants were always removed from protein sequence data sets and one representative gene model was selected instead. This selection was either based on declaration of representative gene models by the responsible annotation groups/consortia or the longest continuous open reading frame (ORF) for a protein-encoding genomic locus. Typically, protein sequences especially from the more fragmented genome assemblies were filtered for internal stop codons and incompatible reading frames as well.

The resulting OrthoMCL gene family output was processed with in-

house scripts to generate statistics on gene group distribution and copy numbers between all species contributing protein sequences. The individual species intersections were represented in three-, four- or five- way VENN diagrams. Moreover, species-specific, lineage-specific, core- and singleton-gene groups were extracted with in-house scripts to facilitate downstream analyses such as GO/PFAM over- and under- representation studies. PFAM domain signatures [142] and GO terms [217] were either derived from SIMAP (SIMAPfeatures instance) [206] or computed with InterproScan 5 (different incremental releases from version 5 beta) [218] using the "-goterms" lookup option. GO terms were computed for all ontology categories while for some analyses only terms from the category "molecular function" were considered because of better transferability between reference and target annotation. The GOstat R package from Bioconductor [219] was used to identify GO terms that appear over- or under-represented in a given dataset (such as expanded gene families) relative to the overall gene set. Significant terms were reported for p-values $=< 0.05$. GOSlim terms were derived for GO terms from the molecular function category only using the AgBase web tool [220] with the 'Plant Slim/TAIR version Aug.2011' GO Slim set. To compute PFAM domains over- or under-represented in a specific sub-sample relative to the overall gene set, in-house software applying Bonferroni correction for multiple testing was used. Expanded gene families were identified in the OrthoMCL output by the species' copy number distribution in a respective gene group using a binomial probability distribution with a significance level < 0.05. For bread wheat, expanded gene families were extracted from (above) the 95% confidence interval of the gene copy frequency distribution whereas contracted gene families were selected from (below) the 5% confidence interval. For the in-depth analysis of particular gene families of interest such as NBS-LRR genes in barley and genes related to hydrogen ion transporter activity in wheat, a combination of BLAST-assisted sequence similarity searches, literature mining, investigation of OrthoMCL gene family clustering results and construction of phylogenetic trees with PROTDIST from the phylip package [221] (bootstrapping with 100 iterations) was used.

2.2 Identification of species- and lineage- specific genes in cereals

To identify both species- and Triticeae- specific transcripts and genes, gene predictions (for barley) and transcriptome assemblies (for wheat) were fil-

tered in a dedicated multi-step analysis pipeline.

For wheat, the input data set of 31,676 repeat-filtered transcript assemblies was processed with this pipeline to identify wheat transcript assemblies sharing homology with publicly available flCDNA sequences from barley [222] only ("Triticeae-specific" transcriptome set), and to identify wheat transcript assemblies specific to wheat and its diploid relatives *Aegilops tauschii* and *Triticum monococcum*. For the Triticeae-specific transcriptome data set a total of 10,088 wheat transcripts was identified with a significant BLASTN[147] hit (evalue < 10e-05) against the barley flCDNA library but with no significant BLASTX hit (evalue < 10e-05) against a comprehensive set of angiosperm reference protein sequences, mainly from finished genome projects including all publicly available grass sequences. For the matched barley flCDNA sequences, GO terms [217] and PFAM domain signatures [142] were extracted from SIMAP [206] and over- and under-represented PFAM and GO terms were computed for this set (see 2.1 for methods). To identify transcripts specifically shared between wheat and its two diploid relatives, repeat-masked transcriptome assemblies were filtered against all orthologous representatives (BLASTX evalue cut-off 10e-05). The resulting 13,345 transcripts without a significant match were searched in the WGS sequences of wheat, *Aegilops tauschii* and *Triticum monococcum* to exclude the possibility of potential contamination in the wheat transcriptome. For 13,103 wheat transcripts, significant BLASTN hits (cut-off 10e-05) against wheat and, optionally, against *Aegilops tauschii* and/or *Triticum monococcum* WGS sequences were reported. Out of these, 439 transcripts only showed significant homology to the wheat WGS sequence but not to *Aegilops tauschii* and *Triticum monococcum* WGS sequences. Last but not least, the remaining 13,103 wheat transcripts were filtered against both the OrthoMCL singleton sequences (all rice, sorghum, *Brachypodium* and barley sequences not found in clusters in the OrthoMCL analysis to define the orthologous representatives) and a comprehensive set of angiosperm reference protein sequences, including all publicly available grass sequences not yet represented in the ortholome data set. In total, 12,604 wheat transcripts were found with no significant BLASTX (cut-off 10e-05) match against at least one of the data sets making them candidates for wheat- specific transcripts or/and transcripts specific to its diploid relatives. Only very few (<10) conserved PFAM domains were identified in this set of transcripts whereas 4,717 of them showed a significant BLASTN (cut-off 10e-05) match against publicly available wheat flCDNA sequences.

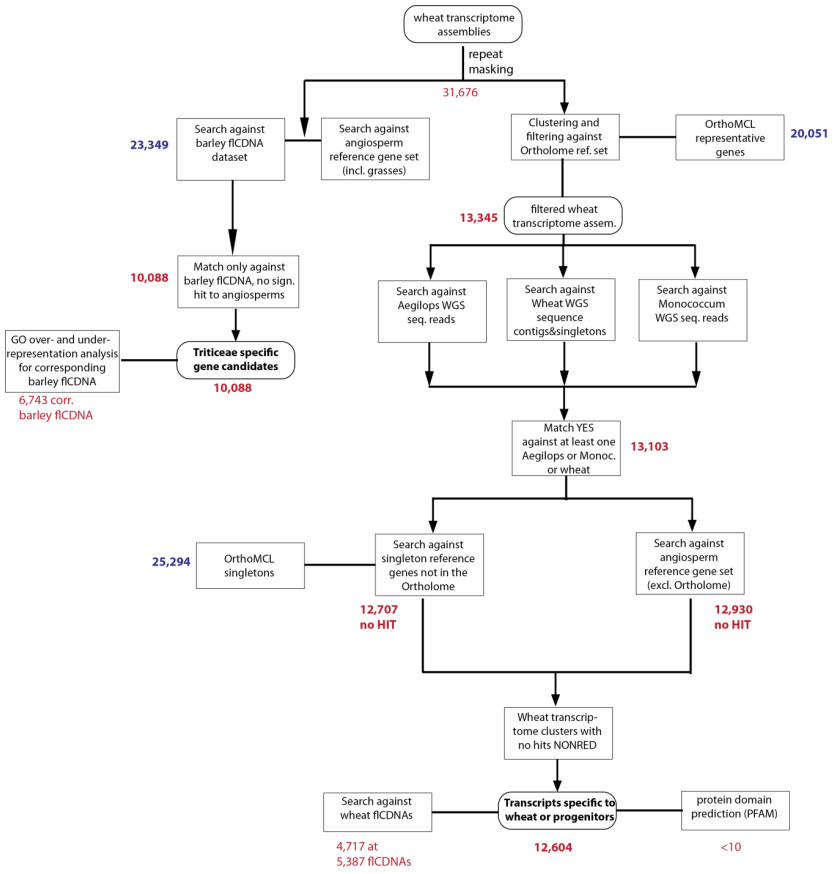

Figure 2.1: Flow chart describing the identification pipeline for Triticeae-specific transcripts.

For barley, all transcripts which were not included in the high-confidence gene set were processed in a pipeline similar to the one outlined for wheat before. Here, a set of 53,220 barley transcripts (49,420 RNA-seq transcripts predictions and 3,800 fl-cDNAs) were scanned for species-specific transcripts as well as nTARs (novel Transcriptional Active Regions) [223] and pseudogenes/remote homologs [2]. In a first step, barley transcripts were identified that share homology with publicly available flCDNA sequences from wheat only ("Triticeae-specific" transcriptome set) as well as transcripts that appear specific to the barley genome. A total of 7,999 transcripts were found with a significant BLASTN hit (e-value $<$ 10e-05) against wheat flCDNA sequences and no significant BLASTX hit (e-value $<$ 10e-05) to a comprehensive set of angiosperm reference protein sequences, mainly from finished genome projects including all publicly available grass sequences.

Barley transcripts that showed no significant homology to any annotated

plant protein sequence and had no match in the NCBI nonRED database (BLAST; e-value cut-off of 10e-05) were searched in the finished genome sequences of rice (MSU_IRGSP_v7 release 31 Oct 2011) and *Brachypodium* (BLASTN; e-value cut-off of 10e-05). A total of 4,830 barley transcripts were found in 13,118 locations on the *Brachypodium* genome and 2,450 barley transcripts matched to 5,844 distinct locations on the rice genome. 2,046 barley transcripts were identified both on the rice genome and on the *Brachypodium* genome, defining the overlap for potentially conserved nTAR elements. When applying more stringent parameters (alignment length >=50% of the originating barley transcript), 282, respectively 124 barley transcripts were found on the *Brachypodium* and rice genome sequences, with a total of 90 barley transcripts identified on both genomes. PFAM domain signatures [142] were computed with InterproScan [218] for the remaining 16,560 barley transcripts that had no homology support in any of the databases compared to. For only 12 distinct barley transcripts PFAM domains could be predicted.

A summary of the analysis pipeline for barley is shown in [2] Supplementary material S7.1.4.

2.3 Classification of gene origin in the hexaploid wheat genome using machine learning

In order to classify gene sub-assemblies generated from the orthologous group assembly in the bread wheat genome to their likely subgenome origin, whole genome sequences were generated for the donor of the wheat D subgenome, *Ae. Tauschii* [224], as well as for the close relative of the A subgenome progenitor *Triticum monococcum* (NCBI archive SRP004490.3), and cDNA sequence assemblies were derived for *Ae. speltoides* (Trick&Bancroft, unpublished data), a member of the *Sitopsis* section to which the putative B genome donor belongs to. "Expecting that A- related sub-assemblies are more related to *T. monococcum* sequences, D- related sub-assemblies to *Ae. tauschii*, and B-related sub-assemblies to *Ae. speltoides*, sequence similarities of the sub-assemblies to each of these datasets would define and discriminate their origin" [225].

BLASTN [147] was used to compute sequence similarities between the wheat sub-assembly sequences and the genomic-, respectively cDNA- sequences from the wheat progenitors. Only sub-assembly genes with matches to all three progenitor sequence pools were subject to downstream classifica-

tion. From the WEKA package [226], different machine learning approaches were tested, trained and applied to discriminate the sequence triplets including Logistic Regression, Naive Bayes, Decision Trees and Support Vector Machine algorithms. As a training set for the machine learning algorithms, sequences from wheat chromosome 1 were used which have been separated into their subgenome origin (A, B and D) using flow cytometry and chromosome sorting before [227]. Performance on the training data set was evaluated by a stratified k-fold cross-validation. Based on these results, a Support Vector Machine (SVM) algorithm from the libSVM package was trained and selected for the final classification and predictions were further filtered by their libSVM probability estimates. From evaluation on the training set, a cut-off of 0.55 was chosen as an optimal trade-off between sample size and accuracy. Consequently, samples with probability estimates smaller than 0.55 were classified as unreliable predictions.

More material can be obtained from [1] Supplementary material section 5.

2.4 PlantsDB: setup of a relational plant genome database system

2.4.1 PlantsDB System Architecture and Design

The PlantsDB genome database framework is implemented in a modular architecture to be able to accommodate new data types and to establish connections between existing and new data entities. As such, the system is composed of different generic data modules which account for the representation of genomic sequence and genome-associated data. PlantsDB's core system is defined by three basic modules: *Clone*, *Contig* and *GeneticElement*. The *Clone* module was set up to hold raw sequence data and all associated information that relates to a physical clone. The *Contigs* module is being used to store all information about the assembly of individual clones into longer contigs and pseudomolecules representing whole chromosomes. The assembled contig sequences are also deposited in the *Contigs* module as well as associations between contigs and super-contigs. Within the third data module, *GeneticElement*, all elements are represented which can be positionally anchored on a genomic sequence ("contigs"). This includes protein encoding genes, regulatory elements such as transcription binding sites, non-coding RNAs, repetitive elements, markers, transposable elements and many

more. To be able to model the hierarchical structure of e.g. a gene model with its exons, introns and UTRs but also splicing variants, *GeneticElement* implements *subelements* and *group* tables allowing the identification of all elements associated with it in simple queries.

All data in PlantsDB is stored in a relational database management system [228]. To facilitate both integration of new functionality and reusability of components, the PlantsDB architecture follows a multi-tier design, consisting of Data tier, Application logic and presentation layer. All middleware components are implemented with the J2EE (Java 2 Enterprise Edition) standard. As an application server, JBOSS release 1.4 [229] is being used and data integration from the data tier is facilitated with JDBC [230]. From middleware to presentation layer, data is communicated in the XML standard [231] and JSP (Java Server Pages) [232] and JSF (Java Server Faces) [233] protocols are used to visualize the data in HTML pages using Cascading style sheets (CSS) [234].

To assist remote access from external tools or applications to PlantsDB data and services, web services were implemented following the standards proposed and developed by the BioMOBY consortium [207, 209].

2.4.2 PlantsDB Analysis Tools, Web Interface and Data Retrieval

Various data formats are supported for the download of sequence data from genetic element reports, including HTML, XML and FASTA format. As a genome browser, the Gbrowse [175] tools has been integrated into the gene reports of many PlantsDB instances. BLAST [147] and its multiple derivatives were integrated into the PlantsDB framework as an homology search engine. The download center of PlantsDB uses the FTP protocol [235] and provides files for download in a magnitude of different formats. Details about the implementation of the CrowsNest Synteny Viewer are given in [3].

Chapter 3

Embedded Publications

This thesis is a cumulative work, incorporating three different first- or corresponding- author publications in peer reviewed journals. Original publications can be obtained from the references given. Additional publications from the author related to topics described or discussed in this thesis can be found in the list of publications.

3.1 Embedded publication 1: Nature 2012 Article - A physical, genetic and functional sequence assembly of the barley genome - The International Barley Genome Sequencing Consortium

Barley is one of the oldest crops domesticated by humans and accounts for an important resource for animal feed, malting and food products nowadays. With a size of ˜5.1 Giga-basepairs and high repeat content, the genome of barley is significantly larger than the human genome and has not been sequenced and analysed up to now as a consequence.

Here, the first whole-genome sequence surveys of the barley varieties *Morex*, *Bowman* and *Barke* are reported, together with a physical map of about 5 Gbp of which 3.9 Gbp could be assigned to genetic positions on barley chromosomes through a high-resolution genetic map constructed from SNV (single-nucleotide variants) and sequence-tag genetic markers.

Analysis of repetitive elements in the barley genome sequence revealed a high proportion of long terminal repeat (LTR) retrotransposons, contributing ˜76% of repeats in random BAC sequences, and a much smaller amount of DNA transposons and non-LTR retrotransposons.

Using RNA-seq expression data from eight different developmental stages, ˜28,000 full-length cDNAs and gene models from related reference organisms mapped to the whole genome sequence assemblies, a total of 26,159 'high-confidence' gene models were identified. These were defined by sequence homology to known reference genes and their presence in multi-organism gene family clusters. Among 53,220 'low-confidence' transcripts not supported by these criteria, potential gene fragments were identified reflecting transposable element activity in cereals. The majority of high-confidence genes could be anchored on the basis of the integrated physical/genetic map framework, revealing higher gene density at the distal ends of the chromosomes. Gene family analysis of barley high-confidence gene predictions with the gene complements of related grass reference organisms identified several gene groups expanded in barley. These include NB-ARC domain proteins related to plant defence responses, (1,3)-β-glucan synthase genes with association to plant–pathogen interactions and sugar transporters and sugar-binding proteins.

Gene expression regulation was studied on RNA-seq data and identified 36-55% of the high-confidence genes to be differentially expressed between

samples from different barley developmental stages, indicating patterns of highly dynamic gene expression and regulation. The identification of extensive alternative splicing (AS) and specific AS regulation in individual samples highlights the significance of post-transcriptional processing which is further supported by many premature termination codon-containing transcripts (PTC+) identified in alternatively spliced genes and the finding of thousands of novel transcriptionally active regions (nTARs) without any homology to protein-coding genes in the barley genome sequence.

Finally, more than 15 million single-nucleotide variants (SNVs) were identified from the sequencing of additional barley cultivars and analysed for their genomic distribution and frequency, shedding light on chromosomal regions associated with low recombination frequency and breeding and domestication signs.

M.S. planned the gene family analysis, contributed to the barley gene prediction (analysis setup, functional descriptors and implementation of the gene confidence classification scheme) and conceived and performed the analysis of non-coding transcriptional active regions in the barley genome. The author also performed all aspects of the gene family analysis in barley. This includes: Identification of barley- and Triticeae-specific genes and gene families; Identification of associated functional categories via over- and under-represented GO/PFAM terms; Comparative gene family analysis with respect to closely related species; Identification of expanded and contracted gene families (copy number variations) with respect to related grass species and biological interpretation of the results. M.S. wrote the corresponding sections for both manuscript and supplemental material. This work is reflected in the sections „Transcribed portion of the barley genome“ (incl. Table 1) and „Regulation of gene expression“ of the main manuscript as well as in multiple sections of Supplemental note 7: S7.1.2 (Figure S15), S7.1.3 (Figure S16, S17), S7.1.4 (Figure S18), S7.2.1, S7.2.6, S7.4.3 (Figure S26).

3.2 Embedded publication 2: Nature 2012 Article - Analysis of the bread wheat genome using whole-genome shotgun sequencing - Rachel Brenchley*, Manuel Spannagl*, Matthias Pfeifer*, Gary L. A. Barker*, Rosalinda D'Amore* et al. *joint first authors

Bread wheat is one of the most important crop plants for human nutrition and grown worldwide under different environments. Up to now, the genome sequence of bread wheat has not been described or analysed, mainly attributed to the large genome size (˜17 Gbp) and complex genetics (allohexaploid with three homeologous subgenomes) involved.

Here, whole genome 454 shotgun sequencing was used to generate a genome survey sequence of *Triticum aestivum* with 5-fold coverage. To prevent highly similar gene copies from the homeologous wheat subgenomes from collapsing in a genome assembly, a strategy was developed which utilizes the finished reference genomes and/or gene predictions from the related grass species *Brachypodium*, rice, sorghum and barley. To assist that, a set of orthologous groups was constructed from the reference genes and one representative gene model (OG representatives) was selected per group. Wheat genomic reads were anchored on each OG and a stringent sub-assembly of the reads gave rise to potential wheat gene models. Assembly parameters were evaluated using simulations with rice and maize datasets to ensure separation of homeologous gene copies in the sub-assembly process. As a result, a total of 94,000-96,000 genes were determined for bread wheat, with thousands of additional gene fragments likely associated with DNA transposons and retroelements. Together with the finding of significant gene loss in gene families this documents the great dynamic in the polyploid wheat genome. Comparison of wheat gene family member distributions to that of the diploid D-subgenome progenitor *Aegilops tauschii* allowed for an estimated gene retention rate of 2.5:1 to 2.7:1 for the modern bread wheat genome after polyploidisation and domestication over ˜10,000 years. Expanded copy numbers were identified for gene families in the wheat genome which are associated to storage proteins, energy production and growth as well as to plant defense mechanisms.

To determine the subgenome origin for each wheat sub-assembly, differing sequence similarities to the projected progenitor genomes of the A,

B and D-subgenomes, namely *Triticum monococcum*, *Aegilops sharonensis/speltoides* and *Aegilops tauschii*, were utilized. About two-thirds of the wheat gene sub-assemblies could be classified after investigating and applying several machine learning approaches. Evaluation with SNP analyses on a subset of the genes confirmed an overall high accuracy of the classification, providing breeders with a new and comprehensive resource for the development of subgenome specific wheat markers. Overall, the classified wheat genic sub-assemblies were categorized into A subgenome (28.3%), B subgenome (29.2%) and D subgenome (33.8%) with the remaining 9% of the assemblies showing internal stop codons .Gene onthology (GO) over- and under-representation analyses of the classified wheat sub-assemblies revealed no significant functional bias in gene loss for any of the subgenomes.

M.S. participated in setting up the analysis strategy and created the orthologous gene set (OG) as well as the functional gene descriptors. I also performed all aspects of the gene family analysis in wheat, this includes: Identification of wheat- and Triticeae-specific genes and gene families; Identification of associated functional categories via over- and under-represented GO/PFAM terms and biological interpretation of the results; Comparative gene family analysis with respect to closely related species; Identification of expanded and contracted gene families (copy number variations) with respect to related grass species and biological interpretation of the results. M.S. developed and implemented the homeologous gene classification strategy and evaluated the machine learning approaches. M.S. wrote the corresponding sections for both manuscript and supplementary material. This work is reflected in the sections „Sequence assembly“, „Genome change in polyploid wheat“ (Figure 3), „Pseudogene analysis“ (Figure 4) and „Determining homeologous relationships of gene assemblies“ of the main manuscript as well as in multiple sections in the Supplementary material: S2.2 (Table S4; Figure S1, S2), S2.3, S2.5, S2.6, S2.9, S2.10, S3.1, S3.2 (Table S10-12; Figure S8, S9), S4.1 (Table S14), S5.1 (Table S16, S17; Figure S11, S12, S13, S14, S15) and S5.3.

3.3 Embedded publication 3: Nucleic Acid Research 2013 - MIPS PlantsDB: a database framework for comparative plant genome research - Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M*. *corresponding author

Since its last description in 2007 in NAR [203], PGSB (formerly MIPS) PlantsDB was actively developed and extended with new data, data types and tools and interfaces. To date, PGSB PlantsDB incorporates more than 20 fully or partially sequenced plant genomes along with multiple associated datasets such as gene predictions, repeat annotation, expression data, metabolic pathways and variation data. PGSB PlantsDB was initially set up as a plant genome database for the emerging whole genome sequence of the first sequenced higher plant *Arabidopsis thaliana* in 2000 [13] and has since evolved towards a comparative platform for the analysis, management and storage of plant genome data.

A major component of PlantsDB is the integration, management and representation of genomic and genetic data from plant reference and model organisms. This includes the genome sequences and annotation of *Arabidopsis thaliana*, *Medicago truncatula*, tomato, barley and wheat. Special focus was given to the integration and representation of genome data from the cereals, attributed to their complex genome structure and resulting data resources. This PlantsDB Triticeae instance includes dedicated views and interfaces for the integrated barley physical and genetic map, the barley genome and gene predictions, the UK wheat gene sub-assemblies (incl. BLAST sequence search) and the visualization, search and download of the GenomeZipper data for barley, wheat, rye and lolium.

Another key component of PGSB PlantsDB is the development of comparative genomics tools to facilitate both knowledge transfer between model and crop species and to address evolutionary questions and assist the analysis of related, more complex genomes. To analyse for and visualize conserved gene order (synteny) between related plant species, the CrowsNest tool was set up and populated with the genomes of *Brachypodium*, sorghum, rice, *Ae. Tauschii* (wheat subgenome progenitor) and barley. Additionally, orthologous genes determined in the OrthoMCL gene family analyses described before were integrated into PlantsDB gene reports for many species,

complemented by fast sequence similarity searches via SIMAP [206].

The MIPS repeat element database (mips-REdat) and catalog (mips-REcat) provide a comprehensive collection of plant repetitive elements together with a classification scheme, facilitating the detection and annotation of transposable elements and repeats in newly sequence plant genomes as well as cross-species comparative analyses.

To enable programmatic data exchange and database crosstalk between distributed plant genome resources in Europe and worldwide, PGSB PlantsDB implements webservices for the remote access to its data and services as developed by the BioMOBY consortium [207]. Moreover, PlantsDB is part of the transPLANT consortium, an European Union initiative to facilitate trans-national infrastructure and inter-connection of plant genome data. Within that project, an international plant genome resource registry is maintained by PlantsDB and cross- search functionality was implemented between major European plant genome resources including PlantsDB, Ensembl plants (EBI), INRA URGI and IPK.

M.S. conceived and implemented the current PlantsDB database infrastructure, integrated and managed the data sets from all species and contributed to the design and implementation of web interfaces and tools (UK454 survey interface; GenomeZipper representation; barley resources; CrowsNest). M.S. wrote and edited the manuscript (including figures and tables) with contributions from all co-authors.

Chapter 4

Discussion

The discussion focuses on five different main aspects related to the work described in chapter 3. In the first parts (chapter 4.1-4.3), the novel findings and results of the gene family analysis performed in this study for the complex Triticeae genomes of barley and wheat (see 3.1 and 3.2) are discussed with respect to previous studies. It will also be shown how the gene and gene family resources established with this thesis have triggered and potentially influence future studies and downstream analyses. The second part (chapter 4.4 and 4.5) of the discussion highlights the new insights into structure and organization of a polyploid Triticeae genome gained with this study for the bread wheat genome (see 3.2) and puts the findings in relation to results from previous analyses in wheat and other polyploid genomes. Chapter 4.5 discusses the implications of the Triticeae genome resources generated in this work (especially gene calls and the separation of homeologous genes in bread wheat) for studies focusing on open questions related to (genome) evolution and domestication of Triticeae. The third part of the discussion (chapter 4.6) focuses on the concepts for the separation and classification of homeologous genes in the bread wheat genome which in the first place allowed for the analyses described in the previous part. Here, alternative approaches for the separation and classification of homeologous genes are introduced and discussed and findings are compared with the results obtained in this study (see 3.1.). Another aspect examined in the fourth part of the discussion (chapter 4.7) focuses on the analysis of transcriptome data in Triticeae genomes. Transcriptome analyses have been performed in the frame of this work for both barley (see 3.2) and wheat (see 3.1). This chapter discusses limitations and benefits of the transcriptome analysis with respect to the

whole genome sequencing approaches discussed previously and illustrates where both can complement each other. Last but not least, implementation and objectives for the efficient integration, management and visualization of complex and heterogenous Triticeae genome data are discussed in part five of the discussion (chapter 4.8). This part relates to the aspects of the Triticeae genome analysis discussed before not only as a dissemination interface but also as a management and integration hub for the complex, heterogenous and often distributed Triticeae genome data.

4.1 Identification of genes and gene families in complex cereal genomes and its implications for crop research and agriculture

The availability of high-coverage genome sequence along with the majority of genes predicted is a big step forward for all researchers working with the cereals wheat and barley. In addition, many barley genes could be positioned on a chromosomal location [2, 236], facilitating target-oriented mapping and isolation of genes underlying trait loci. With that data in hands, genes and gene families can be investigated for cereal-specific characteristics such as gene family expansions or species-specific genes.

With the concepts and novel strategies outlined in this thesis, the complex genomes of the important cereal crops wheat and barley were analyzed for their gene repertoire and gene family composition. Until very recently, no genome-wide studies on gene families in wheat and barley were reported, owing the difficulties involved in genome sequence assembly and analysis of these large and complex Triticeae genomes.

Previous analyses of wheat and barley genes and gene families mainly focused on individual cases relevant or involved in specific biological functions or pathways such as e.g. host-pathogen interactions [91, 92, 237, 238], disease and drought resistance [239-245] and grain yield [246-249]. To quantify the number of specific gene family members and to determine the actual coding sequences of genes in wheat or barley, considerable manpower and laboratory work had to be invested. To support gene analyses, full-length cDNA libraries were constructed for wheat [250] and barley [222]. These libraries, however, are limited to expressed coding sequences and typically remain incomplete with respect to the full gene complement of an organism.

4.2 Comparative analysis of gene families provides new insights into the biology of cereals

With the construction of gene datasets from the whole-genome sequences of wheat and barley - as described in this thesis - , the basis for comprehensive and more advanced gene and gene family analyses was laid. In combination with the availability of numerous genomes from related species such as *Brachypodium* [17], rice [14] and sorghum [4], analytical power increases as many comparative genomics approaches can now also be applied to the genomes of cereals. This has been demonstrated as part of this thesis for both the wheat and barley gene repertoires when OrthoMCL [149] was used to construct orthologous gene families across model and crop cereals. As a result, gene families expanded or reduced in the analyzed cereals were identified and attributed to specific biological functions or traits such as e.g. energy harvesting or disease resistance [1].

For bread wheat, more than 500 reference genes with significantly expanded gene copy numbers relative to its close relative *Aegilops tauschii* were identified ([1] chapter „Genome change in polyploid wheat“ and Supplementary Section 3.2). Functional descriptions associated with these groups include:

- Core histone genes: additional histone gene copies may be advantageous in the formation of heterochromatin for larger and polyploid genomes such as the wheat genome;

- NB-ARC domain containing proteins: these domains have been found to be involved in plant defense response [251] and additional gene copies in the wheat genome may reflect specific defense mechanisms and/or domestication consequences;

- Seed storage proteins account for the majority of the protein content in cereal grains, an indispensable protein resource for human nutrition [252]. As part of the GO functional category “nutrient reservoir activity/storage protein”, several types of storage proteins such as prolamins and plant seed storage proteins were found to be overrepresented in expanded wheat gene families. As storage protein content and variation strongly influences backing quality and other food processing factors in bread wheat and other cereals [253, 254], storage proteins are and have potentially been a major target in breeding and

crop domestication. Both the observed expansion of gene families and variation within storage protein gene members might have been associated with this.

- Photosystem proteins as well as pollen allergens and ribosomal protein components were also found over-represented in both wheat and *Ae. tauschii* expanded gene families providing interesting hints for in-depth investigation of Triticeae gene families and Triticeae biology;

Comparison of bread wheat gene families against its subgenome progenitor *Aegilops tauschii* not only revealed commonly expanded gene families but also dissimilarities which may be attributable to specific species characteristics. As an example, genes encoding hydrogen ion transmembrane transporters (*GO:0015078*) [217] were found in higher copy number in *Ae. tauschii* as compared to bread wheat ([1] Supplementary Section 3.2 and Figure S9). Detailed analysis of these genes revealed their role in the formation of different subunits of ATPases. As a consequence it can be hypothesized that these genes may be involved in providing proton gradients to support Na+ exclusion in *Ae.tauschii* [255] and the accumulation of minerals in other *Aegilops* species [1, 256].

For barley, numerous gene families were identified showing a significantly expanded number of gene members relative to the compared reference organisms rice, *Brachypodium* and sorghum ([2] chapter „Transcribed portion of the barley genome“ and Supplementary Note 7.1.3). The functional descriptions of these gene groups include:

- Expanded copy numbers of sugar-binding proteins and sugar transporters may reflect the prominent storage of carbohydrates in barley grains [257, 258];

- NB-ARC domain gene families known to be involved in plant defense responses [251] were also found to be expanded in the barley genome;

- (1,3)-β-glucan synthase genes were found to be expanded in the barley genome as well, which were previously associated to plant-pathogen interactions [259];

The identification of both expanded/contracted gene families and species-/lineage-specific genes in Triticeae crops provides a jumpboard for in-depth analyses of specific gene families as well as plant characteristics and

underlying traits. On the other hand, findings based on computational methods and observations on gene (family) copy numbers, as outlined above, ask for confirmation and further investigation through experimental approaches before conclusions on their biological relevance or function can be made. Typically, more in-depth studies reporting findings on individual gene families are being published a few years after user communities gain access to the gene repertoires of reference organisms and/or close relatives enabling comparative genome approaches [260-263]. This also highlights the importance and impact of genome-scale studies and the genome data resources developed and made available.

4.3 Gene annotation and construction of gene families in cereals promotes biological studies

The progress in annotating barley and wheat genes and in defining gene families in Triticeae directly promotes insights into biological functions such as plant-pathogen interactions and potentially translates into more optimized and adapted crops in the near future. This is illustrated in a couple of studies published just after the genome sequences and gene calls of wheat and barley became publically available.

Silvar et al. [264] used the integrated physical and genetic map of barley as well as the GenomeZipper to narrow down regions on the genome related to resistance again powdery mildew in barley. Within these regions, candidate genes could be identified from the gene annotation described in this thesis and functional descriptions were queried for terms related to pathogen defense-related processes such as genes encoding proteins from the NBS-LRR class or protein kinase family class.

Muñoz-Amatriaín et al. [265] constructed a comparative genomic hybridization (CGH) array to quantify Copy-number variations (CNVs) in domesticated barley cultivars versus wild barleys. A higher number of CNV diversity was observed in wild barley cultivars. The barley gene predictions and functional annotation were used to investigate the impact of CNV to coding regions. About 9.5% of the coding sequences showed evidence for copy number variations with disease-resistance proteins and protein kinases found to be over-represented in the gene set affected by CNV. This analysis is of particular interest as here CNVs between wild and domesticated cultivars are being investigated on a genome-wide level for a (diploid) Triticeae species for the first time. The results presented in this study may be linked

together with copy number variations observed in gene families constructed between grass model plants and crops, as described in this thesis, to analyse genes and gene families of agronomical relevance.

Kugler et al. [266] used both the wheat LCG assembly [1] and barley high confidence gene predictions to map RNA-seq reads derived from wheat lines either resistant against or susceptible to the *Fusarium* head blight disease. Subsequently, consensus transcripts were constructed for mapped RNA-seq reads and differentially expressed genes could be identified in combination with a network approach for two quantitative trait loci (QTL) related to *Fusarium* head blight resistance. Using functional descriptions and domain annotation, multiple candidate genes and modules, possibly involved in *Fusarium* head blight response in wheat, were extracted. Finally, genes from four different families with known important roles in pathogen response (glucanases, NBS-LRR, WRKY transcription factors and UDP-glycosyltransferases) were extracted and analysed for their expression and pathway location in the expression network.

The results presented in this thesis not only shed new light onto genome characteristics and evolution of two of our major crop plants but also promote insights into the organization and structure of both complex and polyploid plant genomes. They also provide an important basis and foundation for experimental studies on a genome level, not feasible up to now.

4.4 New insights into the structure and organization of complex and polyploid cereal genomes

Pseudogenisation rates in polyploid Triticeae genomes were previously estimated from a small sample of genes. However, the almost full gene complement available now for wheat allows for a refined view on the fate of homeologous gene copies after polyploidisation. In this work it was found that significant gene loss occurred after polyploidisation of the bread wheat genome when compared to its diploid ancestors. With the projected gene retention rate deduced from the comparison of (OrthoMCL) gene family sizes between *Ae. tauschii* and bread wheat, a cumulative loss of 10,000 to 16,000 genes can be concluded in hexaploid wheat ([1] chapter „Genome change in polyploid wheat“ and Supplementary Sections 2.6-2.10 and Figure 3). Previous studies on gene retention and loss in wheat varieties that were formed by recent polyploidisation events report similar ratios [267] accompanied by the massive loss of nucleotide diversity in bread wheat after its

domestication [268]. The analysis of gene fragments identified in high copy number for many wheat genes [1] emphasizes the role of transposable element activity in the de-functionalisation process (pseudogenisation) of gene models and provides un-biased insights into functional domains involved as well as into underlying evolutionary dynamics. Another study highlights the dynamics of pseudogenisation and homeologous gene evolution (alternative splicing patterns, in particular) in hexaploid bread wheat using the flow-sorted genome sequence of wheat chromosome 3A [269], leading to similar conclusions with respect to the consequences of a duplicated genome as well as for the origin of novel traits in the wheat lineage.

No significant preferential gene loss was found for any of the wheat subgenomes, indicating that there is no global subgenome dominance for gene retention and loss ([1] chapter „Determining homeologous relationships of gene assemblies“ and Supplementary Section 5.1 and Figure S13). In addition, no functional categories appeared to be differentially preserved or subject to pseudogenisation and loss in any of the subgenomes ([1] chapter „Determining homeologous relationships of gene assemblies“ and Supplementary Section 5.1 and Figure S14). These findings are in contrast with studies in paleopolyploid maize [18, 270], cotton [271, 272] and soybean [273] where uneven ancient gene loss as well as differentiated gene expression (transcriptional dominance) was reported for the individual subgenomes. This observation may, in part, be explained by the recent polyploidisation of bread wheat when compared to the analysed WGD in maize which was dated to ˜5-12 million years ago [18, 274]. On the contrary, biased gene loss has also been identified for the subgenomes of even more recent polyploids such as *Tragopogon miscellus* [275] (tetraploidisation ˜80 years ago) and synthetic *Arabidopsis* allotetraploids [276], suggesting that different regulatory mechanisms might be involved in the preservation and expression control of subgenome homeologs as well. The upcoming bread wheat reference genome sequence together with comprehensive expression data generated by the IWGSC [95] promises a more detailed view onto this topic.

Studies on the level of sequence variation/similarity between the A-, B- and D-subgenomes of wheat and its progenitor genomes have been based on smaller sub-samples before [81]. Here, an overall high degree of sequence similarity between and within coding sequences from the wheat subgenomes and the progenitor genomes is reported [81]. With almost the full gene repertoire of bread wheat analysed within this study, estimates about the sequence variation of the A-, B- and D-subgenomes of bread wheat was

re-fined, leading to a threshold of 99% identity for the assembly of coding sub-assembly sequences [1].

4.5 The wheat and barley genomes facilitate detailed studies on the evolution and domestication of cereals and their complex genomes

The generation of rich sequence data resources and the decoding of almost all protein-coding genes for barley and wheat now enables in-depth analyses of questions regarding e.g. the evolution and domestication of important cereals. This not only includes the origin of tetraploid and hexaploid wheat and its progenitor genomes but also questions related to the adaption of agricultural important cereal traits and varieties to specific environmental conditions. An example is given by Vigeland et al. [277]. The authors investigate the origin and evolution of genes responsive to low-temperature stimulus in *Pooideae* using substitution rates between orthologous genes from different *Pooideae* species including wheat and barley.

With the sequencing of the close relatives/progenitor genomes of wheat, *Aegilops tauschii* [278] and *Triticum urartu* [279], a basis for genomic studies on wheat domestication and the evolution of a recent allopolyploid cereal is given. Until recently, the analysis of crop domestication (features) with genetic methods could not make use of the full gene repertoire nor the complete genome sequences of important cereals such as wheat and barley. Instead, these approaches had to rely on the analysis of individual gene sets/families or on genome-wide estimates with a few hundred loci using molecular markers (such as amplified fragment length polymorphisms (AFLPs)) to measure genetic similarity [280-282] between geographically separated populations (reviews [283, 284]). It can be expected that comparative genomics studies will benefit from the significantly broadened data basis generated by the whole-genome sequences of wheat and barley. An example might be the use as a mapping reference in NGS re-sequencing projects of cereals.

4.6 Separation and classification of homeologous genes in polyploid cereal genomes

Being able to discriminate and classify homeologous genes in polyploid organisms is of great importance. This is illustrated by the allohexaploid bread

wheat genome with the introduction of specific characteristics by the individual subgenomes. E.g. the wheat D-subgenome particularly contributes to bread backing quality (e.g. via the *Glu-D1d* locus) [285-287] whereas both the A- and B- subgenome holds important genes and QTLs (quantitative trait locus) involved in pathogen defense and resistance [288, 289] (review [290]). However, strong sequence identity between homeologous gene copies may cause merged chimeric assemblies [291] and/or hamper separation and classification of homeologous genes.

In this study, stringent assembly parameters as well as varying sequence similarity of the homeologous (subgenome) genes to the corresponding orthologs in the progenitor genomes were used to discriminate wheat subassemblies and classify them to the A-, B- or D- subgenome. In the classification step both simple sequence similarity e-value/identity threshold-based rules as well as different machine-learning approaches were evaluated. Support Vector Machines (SVM) gave the best compromise between precision and total classified sample size ([1] chapter „Determining homeologous relationships of gene assemblies“ and Supplementary Section 5.1). Evaluation of the results with a SNP-assisted discrimination approach demonstrates overall high precision in separating and classifying homeologous wheat genes using this strategy ([1] chapter „Determining homeologous relationships of gene assemblies“ and Supplementary Section 5.3).

With a significant fraction of the bread wheat genes assembled and assigned to subgenomes, access to the gene repertoire of hexaploid bread wheat is given. This will facilitate the creation of subgenome-specific wheat markers helping in the identification of specific characteristics. It will also support genome-assisted breeding strategies not possible so far in wheat.

Besides whole genome sequencing approaches, transcriptome data can provide a comprehensive overview of an organisms (expressed) gene repertoire (this is discussed in greater detail in section 4.7). When dealing with transcriptome sequencing data (such as ESTs or RNA-seq data) from polyploid organisms, similar limitations and challenges for sequence assembly and separation of homeologous sequences (discussed before) as for genomic data become obvious. Therefore, Schreiber at al. [291] developed a two-stage assembly strategy to generate homeologous-specific assemblies of transcriptome data from hexaploid bread wheat. Both Illumina [22, 23] and Roche 454 [24] technologies were used to generate RNA sequence from multiple developmental stages and tissues. Evaluation of de-novo assemblies generated by different algorithms such as Velvet [27], MIRA [292] and Abyss [28]

on a number of wheat reference genes revealed a high number of chimeric assemblies, collapsing homeologous gene copies. To cope with these problems and also considering the assembly algorithm evaluation results, an assembly pipeline was employed consisting of two different steps: first, the Velvet/Oasis algorithm [27, 293] was used to generate an initial grouping of raw reads into homeologous sequence clusters. These clusters were then further processed individually by applying the MIRA assembly algorithm to generate sequence contigs with high homeolog specificity. Comparing the resulting transcriptome assemblies to a set of publicly available wheat flcDNA sequences, the authors estimate that about 75-80% of the wheat transcriptome should be covered by their assemblies. Homeoelog specificity was determined to be ˜98% based on comparisons against a set of wheat reference genes. For about 70% of the wheat contigs, homeologous sequences were identified in all of the finished reference genomes of *Brachypodium*, rice and sorghum. Among the Gene Ontology (GO) [217] and PFAM [142] descriptions found to be enriched in wheat contigs over rice, RNA and chromatin binding terms as well as translation factor activity terms and transposon-related domains [291] were found. Some of the functions reported here, for example terms related to photosynthesis, were also found in the GO and PFAM over-representation analysis carried out in this study ([1] chapter „Genome change in polyploid wheat“ and Supplementary Section 3.2 and Supplementary Tables 10+11; [2] chapter „Transcribed portion of the barley genome“ and Supplementary Note 7.1.3 and Supplementary Table 25) whereas other categories and terms did not appear to be significantly over- or under-represented here. These differing observations can be attributed to distinct analysis setups. Schreiber et al. considered only the transcribed portion of the genome at defined developmental time points and tissues while the full genomic repertoire of bread wheat was investigated in this study. Schreiber et al. also applied the OrthoMCL software [149] to construct ˜19,000 orthologous gene families from the wheat contigs and the gene predictions from the finished reference species *Brachypodium*, rice and sorghum. As gene groups expanded in wheat, splicing factors and ribonucleoside proteins were identified, a finding consistent with the results obtained in this thesis ([1] Supplementary Table 10+11). Cytochrome P450 proteins were found in contracted gene groups in wheat by Schreiber et al. whereas the PFAM domain signature of these proteins appeared to be over-represented in expanded wheat gene families in the analysis carried out in this thesis ([1] Supplementary Table 11). This observation may be explained by collapsing

paralogous genes with high sequence similarity for this particular gene group in the assembly process used by Schreiber et al.

Both approaches add to the understanding of gene families in cereals while both have individual intrinsic limitations. Overviews about the transcriptome of an organism are always highly dependent on experimental factors such as tissue and developmental stage. As a consequence, transcriptome studies can usually not deliver an overview about the full coding potential of an organisms' genome. On the other hand, strategies as the one outlined in this thesis for wheat, potentially consider genes and coding regions which may no longer be expressed. Another problem faced by transcriptome approaches are alternatively spliced transcripts from the same locus, which may handicap transcriptome assembly and separation of homeologous sequences. Additionally, the classification/association of wheat contigs for/to their specific subgenomic origin remains an unsolved question in the approach described by Schreiber et al.

An alternative approach to discriminate homeologous genes in polyploid (plant) genomes was proposed by Krasileva et al. [90] and applied to the transcriptome of tetraploid emmer wheat ("pasta wheat"). For this the authors first used a multiple k-mer approach to assemble the emmer wheat transcriptome sequences into contigs, demonstrating its advantage over the best single k-mer assembly method. It was shown that this strategy is especially well suited for the de-novo assembly of the transcriptome of tetraploid wheat when compared to the assembly of transcripts from diploid wheat. To separate homeologous wheat genes after the assembly, a phasing approach originally developed for resolving heterozygous haplotypes from next generation sequencing data in humans (HapCUT algorithm [294]) was applied. This strategy involved polymorphism identification, phasing of SNPs, read sorting, and re-assembly of phased reads and resulted in 98.7% correctly separated SNPs from a reference gene data set used for evaluation. In contrast to the separation and classification method outlined in this thesis, the post-assembly phasing pipeline is not dependent on any wheat (or organism)-specific sequence resources making it a sensible method for the separation of transcriptomes from other homozygous tetraploid organisms. As a consequence, however, this strategy does not include a classification step and the association of the separated sub-assemblies to its respective subgenome remains unknown.

Another way to separate and classify homeologous genes in polyploid genomes is chromosome sorting using flow cytometry [295, 296]. Here, the

sequence of individual full chromosomes or chromosome arms is binned and sequenced separately. As a consequence, genes predicted on the binned sequence can clearly be associated to the corresponding subgenome and chromosome. This strategy has been applied by the IWGSC (International Wheat Genome Sequencing Consortium) [94] and generated an even more complete bread wheat gene set and high-confidence classification of homeologous genes. Compared to the approach illustrated for the bread wheat genome in this thesis however, chromosome-sorting involves significantly higher costs and manpower for sequencing and sorting, making it a good choice for establishing high-quality reference sequences.

The sequencing and analysis protocol for complex, polyploid plant genomes developed in this thesis for the bread wheat genome involves an ortholome-directed sub-assembly of exons and separation of homeologs using varying sequence similarities to progenitor/relatives genomes. This approach can be readily applied to other organisms also beyond plants, with large, complex or/and polyploid genomes which have not been sequenced so far because of limitations in dissecting homeologous gene copies. Compared to alternative approaches discussed above this strategy is particularly useful for fast and cost-efficient but comprehensive overviews of the gene repertoire of complex, polyploid genomes. This also includes organisms of economical interest which became polyploid through breeding or domestication such as Triticale, a hybrid between rye (*Secale*) and wheat (*Triticum*) [297].

4.7 Transcriptome data to reveal the expressed portion of cereal genomes

An alternative, although limited, approach to reveal the protein-coding portion of a genome while avoiding the highly repetitive regions is the isolation and sequencing of the transcriptome. This strategy is especially useful when dealing with large and complex genomes such as those from cereals, as demonstrated by Schreiber et al. [291] and discussed above. Several techniques have been established for transcriptome sequencing including RNA-seq [298], cDNA [299] and EST [300] sequencing. Although each method has its own advantages and limitations, RNA-seq sequencing was widely applied lately due to cost efficiency and comparably high sequence yield and depth of sequencing. A limitation common to all transcriptome sequencing methods is that only transcripts expressed under the monitored conditions are reported in the sequence surveys. As a result, comprehensive transcript

libraries need to be constructed for a range of different tissues and conditions to obtain a near-complete overview over the coding potential of a genome [301]. Moreover, associations with genetic data (such as genetic maps) already available for an organism remain complicated and no chromosome ordering or anchoring is given per-se by transcriptome sequence data.

Beside genomic shotgun sequences, transcriptome data were generated for bread wheat from normalized and un-normalized cDNAs using 454 sequencing and used in the analyses carried out in this thesis ([1] Supplementary Section 2.8 and Material & Methods chapter 2.2, Figure 2.1). cDNAs were extracted from three different pools representing several plant tissues and treatments. Wheat transcriptome data were used to determine wheat- and Triticeae-specific transcripts as well as to quantify the transcriptional support for the OG representatives extracted from the genomes of related species (for details see [1] Supplementary Section 2.8 and Material & Methods chapter 2.2, Figure 2.1). Another possible application of the transcriptome data, not investigated within this work, could be in helping to bridge sequence gaps in the wheat gene sub-assemblies, mainly introduced by repetitive DNA in introns.

For barley, the analysis of transcriptome data described in ([2] Supplementary Sections 7.1.4 and following) demonstrated that important mechanisms of gene expression regulation can be studied even on the unfinished and fragmented whole genome sequences of complex cereals. This includes regulation mechanisms mediated by premature termination codons (PTC) ([2] Supplementary Section 7.4) and alternative splicing patterns ([2] Supplementary Sections 7.3).

4.8 Integration, management and visualization of complex genome data within the PlantsDB database framework

One of the key factors determining how efficiently user communities can make use of the newly generated data is its communication through intuitive and comprehensive interfaces. As a result of the novel concepts and strategies applied to analyse the complex and large genomes of cereals, much of the data asks for custom representation and cannot simply be handled in pre-existing genome database infrastructure or tools but asks for dedicated, specifically tailored database solutions. However, powerful genome database

systems are not only needed as a backend for data dissemination but also as a management and integration hub for the complex, heterogenous and often distributed Triticeae genome data generated within this work. As an example, associations between OG representatives in bread wheat (see [1] Supplementary Sections 2.2 and 2.3) and the grass reference genomes were integrated in PlantsDB and used to transfer functional descriptions and sequence features from the model species to wheat ([3] chapter „PLANTSDB-TRITICEAE INSTANCES“).

To assist the tasks of data dissemination, integration and mining for the Triticeae genome data generated, PGSB PlantsDB [3] was extended with a dedicated Triticeae section. Access to data and analysis results is provided in a number of ways:

- Search and browse interfaces provide intuitive access to functional descriptions, structural features of gene predictions and orthologous gene families across model and crop plants;
- Sequence similarity searches using BLAST [147] were implemented to query the barley and wheat gene complements;
- The results from the barley, wheat, rye and Lolium GenomeZippers [64] were integrated and visualized to allow a seamless navigation through individual positions across the chromosomal layout as well as between different positioned genomic features such as markers, ESTs etc.;
- For barley, genetic and physical map data were integrated into a dedicated schema and results are visualized in customized browsers such as GBrowse [175] and CrowsNest;
- Wheat subassembly sequences can be searched and obtained given the corresponding orthologous reference genes from the closely related organisms *Brachypodium*, sorghum, rice and barley;
- Within the CrowsNest tool, regions of conserved gene order (“synteny”) between crop and model plants (e.g. between barley and *Brachypodium*) are pre-computed and visualized in a number of different views and at various zoom levels;
- A download center provides structured access to all bulk data files. Direct access supports local data analyses at the user side;

The intrinsic structure of multi-partner consortia which generated the bulk of data described and used in this work is also reflected in the non-centralized architecture of Triticeae data resources. No single data resource entity is capable of collecting and integrating all data resulting from heterogenous plant genome data sources and developing/providing tools to analyse it. As a consequence, attempts to inter-connect distributed plant genome resources have been started [207, 208] as for example in the frame of the transPLANT project (trans-National Infrastructure for Plant Genomic Science) [302]. Different strategies and technologies have been proposed to aggregate data and information from distributed resources (see introduction). The benefit for the end user can be seen in a virtually integrated search interface or result page, hiding the underlying complexity and segmentation of data and data types. However, for the time being, no single technology or protocol can be identified as a standard for establishing interconnections between genome databases. As a result, the implementation, efficiency and dimension of cross-talk between plant genome databases is dependent and influenced by dedicated projects among partners. It can be expected, however, that this strategy/approach will be strengthened and become even more popular with the emerging use of semantic web ("Web3.0") technologies [303] in database interfaces and visualisation tools.

Chapter 5

Outlook

5.1 Gene and gene family analysis benefits from finished grass genome sequences

Although the almost complete gene repertoires for barley and wheat were identified with the strategies outlined in this thesis, only a part (barley) or none of the genes (wheat) could be directly assigned to a concrete genomic location or/and ordered along the chromosomes with respect to its neighboring genes. However, both the precise localization of genes on chromosomes and the ordering of genetic elements in a distinct genetic region is of importance for the development of genetic markers and the identification of loci responsible for specific trait characteristics such as disease resistance [348].

In addition, the further analysis of the gene repertoire and gene families in cereal genomes would greatly benefit from finished chromosome sequences (or from at least ordered, contiguous sequence scaffolds). In detail, some of the aspects include:

a.) positionally anchored and ordered gene models on a contiguous sequence allow for the identification of tandem or/and segmental gene duplications [119]. These are important mechanisms contributing to gene family expansions and being able to understand the mode of a gene family expansion may help to shed light on their evolution and biological role [137-140].

b.) positionally anchored and ordered gene models on a contiguous sequence also facilitate studies on the synteny (conserved gene order; see chapter 1.2.4) [69] between gene family members and between orthologous genes. Analysing expanded gene families or gene families with species-specific gene loss, the syntenic context can provide valueable information about chro-

mosomal regions of high sequence conservation or divergence. Taking the syntenic context into account also helps in the identification of orthologous genes between closely related species, especially in the presence of highly similar paralogous genes, which is a prominent problem in the complex grass genomes.

c.) positionally anchored and ordered gene models on a contiguous sequence also facilitate fine-grained studies on gene borders, exon/intron boundaries, transcription-factor binding sites, neighboring or intersecting transposable elements (TE) and repeat elements and many other elements [379]. This becomes especially relevant and interesting when studying the evolution of specific gene families such as storage proteins in bread wheat or resistance genes in barley.

5.2 High-quality reference genome sequences are mandatory for many genome-scale analyses

Genome-wide association studies (GWAS) (reviews [349, 350]) as well as (population) re-sequencing studies involving thousands of different varieties or individuals [351, 352] have demonstrated their value for the identification of loci associated with diseases [353, 354] or specific trait characteristics [355, 356] in a broad spectrum of species. GWAS and re-sequencing approaches do not require finished reference genome sequences to map the generated NGS reads against, however, in practice, absence of a high-quality reference genome complicates the study design and analysis of results [357, 358], especially when dealing with polyploid and large genomes. As a result, GWAS studies reported for plants focused on species with finished genome sequences so far (reviews [359-361]) [356, 362]. Re-sequencing projects with the intention to gain new insights into population genetics/genomics and evolution have been started for the model plant *Arabidopsis thaliana* (1001 *Arabidopsis* project [12, 352]) as well as for rice (review [363]) [364], maize (review [365]) [366-368] and numerous projects have been initiated for additional plant species including barley and wheat.

Other methods to discriminate varieties and determine polymorphisms and genetic variation in plants such as DNA profiling and fingerprinting [369, 370] also benefit from finished genome sequences as marker and primer development often targets non-protein coding regions such as simple sequence repeats (SSR) [371].

For repetitive elements many studies highlighted their role in the evo-

lution of epigenetic regulation as well as their long-term impact on genome stability and evolution [34, 372]. In the assembly of NGS sequences from large and complex genomes such as cereal genomes, repetitive elements are often collapsed into single number entities due to their high sequence similarity. As a result, estimates for the transposable element content and repeat composition in large genomes have to be based on short read and/or k-mer frequencies and TE/gene reference templates, as demonstrated for the genomes of barley [2] and wheat [1]. Consequently, more detailed studies on the repeat content and the impact of repetitive elements on regulation and genome stability in the large and complex Triticeae genomes would greatly benefit from finished reference genomes.

5.3 Beyond gene annotation and expression – regulation and epigenetic mechanisms to control grass phenotypes

Besides genes and their products, additional layers of information contribute to the expression of biological functions and to phenotypes. The "second layer" - regulation of genes and gene expression by a plethora of mechanisms [304-306] – so far was not thoroughly studied for complex cereal genomes. One main reason are difficulties involved in generating longer continuous sequence scaffolds for such large genomes in the presence of high repeat content. As described before, even advanced and specialized algorithms cannot overcome the problems of assembling short sequence reads into long continuous scaffolds in a high repeat content background [99]. Especially analyses related to the identification of (gene) regulatory elements and mechanisms would benefit most from the availability of finished pseudo-chromosome or at least long scaffold genome sequences [307, 308]. This includes:

a.) currently, extraction of sequences flanking coding regions in barley and wheat is limited. Typically, a region of around 1000 base-pairs upstream of the predicted transcription start site of a gene is used to search for transcription factor binding sites [309-311]. This region is also often called a putative promoter region and a number of different methods and strategies can be applied to identify novel or conserved gene regulatory element binding sites [307, 312-314]. Methods such as phylogenetic footprinting [315, 316] with its implementations FootPrinter [317] and ConSite/Phylofoot [318] search for conserved motifs in the promoter sequences of orthologous genes

from multiple related organisms. Regulatory elements which appear well conserved over many species can also by identified by probabilistic frameworks [319], patterns/motifs [320] or Markov models [321, 322] and universal frameworks such as FIRE [323] and GEMS [324] incorporate additional components such as (co-) expression measurements.

b.) MicroRNAs are small (˜21-22 nucleotides), non-coding RNAs regulating the gene expression by base-pairing with complementary sequences. This consequently leads to the silencing of a gene either via translational repression or target degradation [325]. MicroRNAs were identified in flies [326], worms [327], mammals [328] and plants [329, 330], with currently ˜300 different microRNAs annotated for *Arabidopsis thaliana* in miRBase [331]. In plants, microRNAs are primarily involved in regulation of growth and development [332] and many plant microRNAs appear to be evolutionary conserved even between more distantly related species such as *Arabidopsis thaliana* and rice [333]. As both microRNA genes and targets are encoded in nuclear DNA they can be studied best on finished whole genome sequences. Although some estimates about microRNAs such as copy numbers and presence/absence can be derived even from fragmented and unfinished plant genome sequences more thorough studies such as prediction of microRNA targets [334], precursors [335] and microRNA-like structures [336] require finished genome sequences or longer sequence scaffolds [337].

c.) Expression and accessibility of particular genes and/or chromosomal regions is often controlled by epigenetic regulation mechanisms such as DNA methylation and histone modifications [338-341]. These mechanisms act on chromatin structure and genome stability, making their understanding indispensable for the "proper interpretation of genetic information and determination of phenotypes" [342]. Methylation of DNA is one of the most prominent epigenetic mechanisms to either temporarily or permanently silence gene expression [343]. Here, a methyl group is being added to a cytosine nucleotide, resulting in significantly reduced accessibility of DNA [344]. In plants, DNA methylation was found to particularly affect transposable elements and other repetitive DNA elements [342, 345, 346]. Genome-wide methylation profiles (or "methylomes") have been determined and analysed at single-basepair-resolutions for some plant species including *Arabidopsis thaliana* (two ecotypes and their reciprocal hybrids) [342, 347] but not for the more complex cereal genomes with fragmented and yet unfinished genome sequences.

5.4 Towards contiguous chromosome sequences for the complex cereals wheat and barley

The results and resources derived from NGS sequencing and analysing the barley and bread wheat genomes shed new light on the biology of these important cereals and provide an important foundation of tremendous value for upcoming studies and agricultural applications. However, as outlined above, some analyses and studies require continuous, ordered and/or finished genome sequences. This is also especially relevant for the isolation of individual genes and for joining genetic data such as markers with the genomic data generated. As a consequence, work on longer sequence scaffolds and/or finished pseudo-chromosome sequences for barley and bread wheat are on-going efforts, even though this is expected to be a time- and resource-intensive process due to complexity and size of the cereal genomes. These efforts include:

a.) Chromosome-sorting of bread wheat NGS sequences using flow cytometry [295, 296] and subsequent assembly of individual chromosome (arm) sequences within the IWGSC (International Wheat Sequencing Consortium) [94, 95]. Using this strategy, the problems involved with high-dimensional sequence repeats and unknown chromosome association can be addressed and reduced (although not solved). This approach is expected to yield a high-confidence separation of homeologous genes for the close-to-complete gene repertoire of hexaploid bread wheat.

b.) Sequencing of the minimum tiling path (MTP) [373] in barley [374]. Within the International Barley Sequencing Consortium (IBSC) funding and efforts are coordinated to individually sequence the minimum tiling path for all barley chromosomes [93]. This will, in the best scenario, result in a high quality reference genome sequence with pseudo-chromosome sequences or/and longer sequence scaffolds and enable access to the full gene complement. It will also allow detailed studies on genome structure (repeats, regulatory elements) and organization.

Sequencing technologies rapidly evolved over the last couple of years and, as a result, sequencing costs significantly decreased. For the near future, further technical advances can be expected [97, 375]. This includes the availability of longer sequence reads generated by advanced NGS technology. Whereas current NGS reads usually consist of 50-700 bp (e.g. Illumina [23] and Roche 454 [24] platforms), upcoming or already introduced sequencing platforms such as from Pacific Biosciences [376] and Oxford Nanopore [377]

promise to deliver sequence reads of up to 10,000 bp and more in length. With sequence reads of this length, stretches of repetitive genome DNA can be spanned more easily when assembling contigs and scaffolds from the raw sequence reads. Sequencing and/or improving the sequence of the large and complex cereal genomes represents an exciting and promising application for these technologies.

Chapter 6

References

1. Brenchley, R., et al., *Analysis of the bread wheat genome using whole-genome shotgun sequencing.* Nature, 2012. **491**(7426): p. 705-10.

2. International Barley Genome Sequencing, C., et al., *A physical, genetic and functional sequence assembly of the barley genome.* Nature, 2012. **491**(7426): p. 711-6.

3. Nussbaumer, T., et al., *MIPS PlantsDB: a database framework for comparative plant genome research.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1144-51.

4. Paterson, A.H., et al., *The Sorghum bicolor genome and the diversification of grasses.* Nature, 2009. **457**(7229): p. 551-6.

5. Huang, S., et al., *The genome of the cucumber, Cucumis sativus L.* Nat Genet, 2009. **41**(12): p. 1275-81.

6. Wang, K., et al., *The draft genome of a diploid cotton Gossypium raimondii.* Nat Genet, 2012. **44**(10): p. 1098-103.

7. Koboldt, D.C., et al., *The next-generation sequencing revolution and its impact on genomics.* Cell, 2013. **155**(1): p. 27-38.

8. Michelmore, R., *Genomic approaches to plant disease resistance.* Curr Opin Plant Biol, 2000. **3**(2): p. 125-31.

9. Mayer, K.F., et al., *Role of WUSCHEL in regulating stem cell fate in the Arabidopsis shoot meristem.* Cell, 1998. **95**(6): p. 805-15.

10. Proost, S., et al., *PLAZA: a comparative genomics resource to study gene and genome evolution in plants.* Plant Cell, 2009. **21**(12): p. 3718-31.

11. Paterson, A.H., et al., *Comparative genomics of plant chromosomes.* Plant Cell, 2000. **12**(9): p. 1523-40.

12. Cao, J., et al., *Whole-genome sequencing of multiple Arabidopsis*

thaliana populations. Nat Genet, 2011. **43**(10): p. 956-63.

13. Arabidopsis Genome, I., *Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.* Nature, 2000. **408**(6814): p. 796-815.

14. International Rice Genome Sequencing, P., *The map-based sequence of the rice genome.* Nature, 2005. **436**(7052): p. 793-800.

15. Tomato Genome, C., *The tomato genome sequence provides insights into fleshy fruit evolution.* Nature, 2012. **485**(7400): p. 635-41.

16. Young, N.D., et al., *The Medicago genome provides insight into the evolution of rhizobial symbioses.* Nature, 2011. **480**(7378): p. 520-4.

17. International Brachypodium, I., *Genome sequencing and analysis of the model grass Brachypodium distachyon.* Nature, 2010. **463**(7282): p. 763-8.

18. Schnable, P.S., et al., *The B73 maize genome: complexity, diversity, and dynamics.* Science, 2009. **326**(5956): p. 1112-5.

19. Pellicer, J., M.F. Fay, and I.J. Leitch, *The largest eukaryotic genome of them all?* Botanical Journal of the Linnean Society, 2010. **164**(1): p. 10-15.

20. Garnatje, T., et al., *GSAD: a genome size in the Asteraceae database.* Cytometry A, 2011. **79**(6): p. 401-4.

21. Venter, J.C., et al., *The sequence of the human genome.* Science, 2001. **291**(5507): p. 1304-51.

22. Bentley, D.R., et al., *Accurate whole human genome sequencing using reversible terminator chemistry.* Nature, 2008. **456**(7218): p. 53-59.

23. *Illumina Next-Generation Sequencing.* [cited 2014; Available from: http://www.illumina.com/.

24. Rothberg, J.M. and J.H. Leamon, *The development and impact of 454 sequencing.* Nat Biotechnol, 2008. **26**(10): p. 1117-24.

25. Morey, M., et al., *A glimpse into past, present, and future DNA sequencing.* Molecular Genetics and Metabolism, 2013. **110**(1-2): p. 3-24.

26. Liu, L., et al., *Comparison of next-generation sequencing systems.* J Biomed Biotechnol, 2012. **2012**: p. 251364.

27. Zerbino, D.R. and E. Birney, *Velvet: algorithms for de novo short read assembly using de Bruijn graphs.* Genome Res, 2008. **18**(5): p. 821-9.

28. Simpson, J.T., et al., *ABySS: a parallel assembler for short read sequence data.* Genome Res, 2009. **19**(6): p. 1117-23.

29. Margulies, M., et al., *Genome sequencing in microfabricated high-density picolitre reactors.* Nature, 2005. **437**(7057): p. 376-80.

30. Butler, J., et al., *ALLPATHS: de novo assembly of whole-genome*

shotgun microreads. Genome Res, 2008. **18**(5): p. 810-20.

31. Scheibye-Alsing, K., et al., *Sequence assembly.* Comput Biol Chem, 2009. **33**(2): p. 121-36.

32. Baucom, R.S., et al., *Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome.* PLoS Genet, 2009. **5**(11): p. e1000732.

33. Wicker, T., et al., *Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives.* Plant Cell, 2011. **23**(5): p. 1706-18.

34. Jurka, J., et al., *Repetitive sequences in complex genomes: Structure and evolution.* Annual Review of Genomics and Human Genetics, 2007. **8**: p. 241-259.

35. Salzberg, S.L. and J.A. Yorke, *Beware of mis-assembled genomes.* Bioinformatics, 2005. **21**(24): p. 4320-1.

36. Semon, M. and K.H. Wolfe, *Consequences of genome duplication.* Curr Opin Genet Dev, 2007. **17**(6): p. 505-12.

37. Adams, K.L. and J.F. Wendel, *Polyploidy and genome evolution in plants.* Curr Opin Plant Biol, 2005. **8**(2): p. 135-41.

38. Jiao, Y., et al., *Ancestral polyploidy in seed plants and angiosperms.* Nature, 2011. **473**(7345): p. 97-100.

39. Schmutz, J., et al., *Genome sequence of the palaeopolyploid soybean.* Nature, 2010. **463**(7278): p. 178-83.

40. Bolot, S., et al., *The 'inner circle' of the cereal genomes.* Curr Opin Plant Biol, 2009. **12**(2): p. 119-25.

41. Blanc, G. and K.H. Wolfe, *Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes.* Plant Cell, 2004. **16**(7): p. 1667-78.

42. Hufton, A.L. and G. Panopoulou, *Polyploidy and genome restructuring: a variety of outcomes.* Curr Opin Genet Dev, 2009. **19**(6): p. 600-6.

43. Byrne, K.P. and K.H. Wolfe, *Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication.* Genetics, 2007. **175**(3): p. 1341-50.

44. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.

45. Comai, L., *The advantages and disadvantages of being polyploid.* Nat Rev Genet, 2005. **6**(11): p. 836-46.

46. Wood, T.E., et al., *The frequency of polyploid speciation in vascular*

plants. Proc Natl Acad Sci U S A, 2009. **106**(33): p. 13875-9.

47. D'Hont, A., et al., *The banana (Musa acuminata) genome and the evolution of monocotyledonous plants.* Nature, 2012. **488**(7410): p. 213-7.

48. Blekhman, R., A. Oshlack, and Y. Gilad, *Segmental Duplications Contribute to Gene Expression Differences Between Humans and Chimpanzees.* Genetics, 2009. **182**(2): p. 627-630.

49. Gu, Z.L., et al., *Duplicate genes increase gene expression diversity within and between species.* Nature Genetics, 2004. **36**(6): p. 577-579.

50. Zhang, P., S. Chopra, and T. Peterson, *A segmental gene duplication generated differentially expressed myb-homologous genes in maize.* Plant Cell, 2000. **12**(12): p. 2311-2322.

51. Yanai, Y., et al., *Genomic organization of 251 kDa acetyl-CoA carboxylase genes in Arabidopsis: tandem gene duplication has made two differentially expressed isozymes.* Plant Cell Physiol, 1995. **36**(5): p. 779-87.

52. Van de Peer, Y., *Computational approaches to unveiling ancient genome duplications.* Nat Rev Genet, 2004. **5**(10): p. 752-63.

53. Wendel, J.F., *Genome evolution in polyploids.* Plant Mol Biol, 2000. **42**(1): p. 225-49.

54. Seoighe, C., *Turning the clock back on ancient genome duplication.* Curr Opin Genet Dev, 2003. **13**(6): p. 636-43.

55. Potato Genome Sequencing, C., et al., *Genome sequence and analysis of the tuber crop potato.* Nature, 2011. **475**(7355): p. 189-95.

56. Tang, H., et al., *Synteny and collinearity in plant genomes.* Science, 2008. **320**(5875): p. 486-8.

57. Paterson, A.H., et al., *Comparative genomics of grasses promises a bountiful harvest.* Plant Physiol, 2009. **149**(1): p. 125-31.

58. Chen, M., P. SanMiguel, and J.L. Bennetzen, *Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice.* Genetics, 1998. **148**(1): p. 435-43.

59. Tikhonov, A.P., et al., *Colinearity and its exceptions in orthologous adh regions of maize and sorghum.* Proc Natl Acad Sci U S A, 1999. **96**(13): p. 7409-14.

60. Song, R., V. Llaca, and J. Messing, *Mosaic organization of orthologous sequences in grass genomes.* Genome Res, 2002. **12**(10): p. 1549-55.

61. Ling, H.Q., Y. Zhu, and B. Keller, *High-resolution mapping of the leaf rust disease resistance gene Lr1 in wheat and characterization of BAC clones from the Lr1 locus.* Theor Appl Genet, 2003. **106**(5): p. 875-82.

62. Yan, L.L., et al., *The wheat VRN2 gene is a flowering repressor*

down-regulated by vernalization. Science, 2004. **303**(5664): p. 1640-1644.

63. Mayer, K.F., et al., *Gene content and virtual gene order of barley chromosome 1H.* Plant Physiol, 2009. **151**(2): p. 496-505.

64. Mayer, K.F., et al., *Unlocking the barley genome by chromosomal and comparative genomics.* Plant Cell, 2011. **23**(4): p. 1249-63.

65. Hulbert, S.H., et al., *Genetic mapping and characterization of sorghum and related crops by means of maize DNA probes.* Proc Natl Acad Sci U S A, 1990. **87**(11): p. 4251-5.

66. Bennetzen, J.L. and M. Freeling, *The unified grass genome: synergy in synteny.* Genome Res, 1997. **7**(4): p. 301-6.

67. Devos, K.M. and M.D. Gale, *Comparative genetics in the grasses.* Plant Mol Biol, 1997. **35**(1-2): p. 3-15.

68. Gale, M.D. and K.M. Devos, *Comparative genetics in the grasses.* Proc Natl Acad Sci U S A, 1998. **95**(5): p. 1971-4.

69. Bennetzen, J.L., *Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions.* Plant Cell, 2000. **12**(7): p. 1021-9.

70. Feuillet, C. and B. Keller, *High gene density is conserved at syntenic loci of small and large grass genomes.* Proc Natl Acad Sci U S A, 1999. **96**(14): p. 8265-70.

71. Moore, G., et al., *Cereal genome evolution. Grasses, line up and form a circle.* Curr Biol, 1995. **5**(7): p. 737-9.

72. Devos, K.M., et al., *Chromosomal rearrangements in the rye genome relative to that of wheat.* Theor Appl Genet, 1993. **85**(6-7): p. 673-80.

73. Paterson, A.H., et al., *Toward a unified genetic map of higher plants, transcending the monocot-dicot divergence.* Nat Genet, 1996. **14**(4): p. 380-2.

74. Salse, J., *In silico archeogenomics unveils modern plant genome organisation, regulation and evolution.* Curr Opin Plant Biol, 2012. **15**(2): p. 122-30.

75. Purugganan, M.D. and D.Q. Fuller, *The nature of selection during plant domestication.* Nature, 2009. **457**(7231): p. 843-8.

76. Harlan, J.R. and D. Zohary, *Distribution of wild wheats and barley.* Science, 1966. **153**(3740): p. 1074-80.

77. Dvorak, J. and E.D. Akhunov, *Tempos of gene locus deletions and duplications and their relationship to recombination rate during diploid and polyploid evolution in the Aegilops-Triticum alliance.* Genetics, 2005.

171(1): p. 323-32.

78. Eckardt, N.A., *Evolution of domesticated bread wheat.* Plant Cell, 2010. **22**(4): p. 993.

79. Dvorak, J., et al., *Molecular characterization of a diagnostic DNA marker for domesticated tetraploid wheat provides evidence for gene flow from wild tetraploid wheat to hexaploid wheat.* Mol Biol Evol, 2006. **23**(7): p. 1386-96.

80. Paux, E., et al., *A physical map of the 1-gigabase bread wheat chromosome 3B.* Science, 2008. **322**(5898): p. 101-4.

81. Petersen, G., et al., *Phylogenetic relationships of Triticum and Aegilops and evidence for the origin of the A, B, and D genomes of common wheat (Triticum aestivum).* Mol Phylogenet Evol, 2006. **39**(1): p. 70-82.

82. Golovnina, K.A., et al., *[Phylogeny of the A genomes of wild and cultivated wheat species].* Genetika, 2009. **45**(11): p. 1540-7.

83. Allaby, R.G. and T.A. Brown, *Identification of a 5S rDNA spacer type specific Triticum urartu and wheats containing the T. urartu genome.* Genome, 2000. **43**(2): p. 250-4.

84. Salse, J., et al., *New insights into the origin of the B genome of hexaploid wheat: evolutionary relationships at the SPA genomic region with the S genome of the diploid relative Aegilops speltoides.* BMC Genomics, 2008. **9**: p. 555.

85. Blake, N.K., et al., *Phylogenetic reconstruction based on low copy DNA sequence data in an allopolyploid: the B genome of wheat.* Genome, 1999. **42**(2): p. 351-60.

86. Haider, N., *The origin of the B-genome of bread wheat (Triticum aestivum L.).* Genetika, 2013. **49**(3): p. 303-14.

87. Dvorak, J., et al., *The structure of the Aegilops tauschii genepool and the evolution of hexaploid wheat.* Theoretical and Applied Genetics, 1998. **97**(4): p. 657-670.

88. Matsuoka, Y. and S. Nasuda, *Durum wheat as a candidate for the unknown female progenitor of bread wheat: an empirical study with a highly fertile F-1 hybrid with Aegilops tauschii Coss.* Theoretical and Applied Genetics, 2004. **109**(8): p. 1710-1717.

89. Marcussen, T., et al., *Ancient hybridizations among the ancestral genomes of bread wheat.* Science, 2014. **345**(6194): p. 1250092.

90. Krasileva, K.V., et al., *Separating homeologs by phasing in the tetraploid wheat transcriptome.* Genome Biol, 2013. **14**(6): p. R66.

91. Afanasenko, O., et al., *Genetics of host-pathogen interactions in the*

Pyrenophora teres f. teres (net form) - barley (Hordeum vulgare) pathosystem. European Journal of Plant Pathology, 2007. **117**(3): p. 267-280.

92. Ghazvini, H. and A. Tekauz, *Host-pathogen interactions among barley genotypes and Bipolaris sorokiniana isolates.* Plant Disease, 2008. **92**(2): p. 225-233.

93. Schulte, D., et al., *The international barley sequencing consortium–at the threshold of efficient access to the barley genome.* Plant Physiol, 2009. **149**(1): p. 142-7.

94. Gill, B.S., et al., *A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium.* Genetics, 2004. **168**(2): p. 1087-96.

95. Moolhuijzen, P., et al., *Wheat genome structure and function: genome sequence data and the International Wheat Genome Sequencing Consortium.* Australian Journal of Agricultural Research, 2007. **58**(6): p. 470-475.

96. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.

97. Mardis, E.R., *Next-generation sequencing platforms.* Annu Rev Anal Chem (Palo Alto Calif), 2013. **6**: p. 287-303.

98. Miller, J.R., S. Koren, and G. Sutton, *Assembly algorithms for next-generation sequencing data.* Genomics, 2010. **95**(6): p. 315-27.

99. Schatz, M.C., A.L. Delcher, and S.L. Salzberg, *Assembly of large genomes using second-generation sequencing.* Genome Research, 2010. **20**(9): p. 1165-1173.

100. Poland, J.A., et al., *Development of High-Density Genetic Maps for Barley and Wheat Using a Novel Two-Enzyme Genotyping-by-Sequencing Approach.* Plos One, 2012. **7**(2).

101. Spannagl, M., *Triangulare vergleichende Genomanalyse pflanzlicher Genome: genomische Dynamik evolutionär verwandter Arten*, in *Institut für Genomorientierte Bioinformatik der Technischen Universität München.* 2009, Technische Universität München/Ludwig-Maximilians-Universität München: München.

102. Chaw, S.M., et al., *Dating the monocot-dicot divergence and the origin of core eudicots using whole chloroplast genomes.* Journal of Molecular Evolution, 2004. **58**(4): p. 424-441.

103. Kellogg, E.A., *Evolutionary history of the grasses.* Plant Physiol, 2001. **125**(3): p. 1198-205.

104. Jacobs, B.F., J.D. Kingston, and L.L. Jacobs, *The origin of grass-*

dominated ecosystems. Annals of the Missouri Botanical Garden, 1999. **86**(2): p. 590-643.

105. Barker, N.P., et al., *Phylogeny and subfamilial classification of the grasses (Poaceae).* Annals of the Missouri Botanical Garden, 2001. **88**(3): p. 373-457.

106. Hsiao, C., et al., *A molecular phylogeny of the grass family (Poaceae) based on the sequences of nuclear ribosomal DNA (ITS).* Australian Systematic Botany, 1999. **11**(5-6): p. 667-688.

107. Gaut, B.S., *Evolutionary dynamics of grass genomes.* New Phytologist, 2002. **154**(1): p. 15-28.

108. Hancock, J.F., *Contributions of domesticated plant studies to our understanding of plant evolution.* Annals of Botany, 2005. **96**(6): p. 953-963.

109. Diamond, J., *Evolution, consequences and future of plant and animal domestication.* Nature, 2002. **418**(6898): p. 700-7.

110. Sweeney, M. and S. McCouch, *The complex history of the domestication of rice.* Ann Bot, 2007. **100**(5): p. 951-7.

111. Pingali, P.L., *Green revolution: impacts, limits, and the path ahead.* Proc Natl Acad Sci U S A, 2012. **109**(31): p. 12302-8.

112. *Food and Agriculture Organisation of the United Nations FAOSTAT.* 2014.

113. *FAOSTAT Food and Agricultural commodities production for 2012 (Rankings).* 2012; Available from: http://faostat3.fao.org/faostat-gateway/go/to/browse/rankings/commodities_by_regions/E.

114. Blake, T., Blake, V., Bowman, J. & Abdel-Haleem, H., *Barley: Production, Improvement and Uses* S.E. Ullrich, Editor. 2011. p. 522–531

115. Nevo, E., et al., *Evolution of wild cereals during 28 years of global warming in Israel.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(9): p. 3412-3415.

116. Grando, S.M., H. G. . *Proceedings of the International Workshop on Food Barley Improvement.* 14–17 January 2002. Hammamet, Tunisia.

117. Flagel, L.E. and J.F. Wendel, *Gene duplication and evolutionary novelty in plants.* New Phytologist, 2009. **183**(3): p. 557-564.

118. Innan, H. and F. Kondrashov, *The evolution of gene duplications: classifying and distinguishing between models.* Nature Reviews Genetics, 2010. **11**(2): p. 97-108.

119. Bennetzen, J.L., *Transposable elements, gene creation and genome rearrangement in flowering plants.* Curr Opin Genet Dev, 2005. **15**(6): p.

621-7.

120. Keeling, P.J. and J.D. Palmer, *Horizontal gene transfer in eukaryotic evolution.* Nat Rev Genet, 2008. **9**(8): p. 605-18.

121. Raymond, J. and R.E. Blankenship, *Horizontal gene transfer in eukaryotic algal evolution.* Proc Natl Acad Sci U S A, 2003. **100**(13): p. 7419-20.

122. Sankoff, D., C. Zheng, and Q. Zhu, *The collapse of gene complement following whole genome duplication.* BMC Genomics, 2010. **11**: p. 313.

123. Lai, J., et al., *Gene loss and movement in the maize genome.* Genome Res, 2004. **14**(10A): p. 1924-31.

124. Emes, R.D., et al., *Comparison of the genomes of human and mouse lays the foundation of genome zoology.* Hum Mol Genet, 2003. **12**(7): p. 701-9.

125. Hu, T.T., et al., *The Arabidopsis lyrata genome sequence and the basis of rapid genome size change.* Nat Genet, 2011. **43**(5): p. 476-81.

126. Marino-Ramirez, L., I.K. Jordan, and D. Landsman, *Multiple independent evolutionary solutions to core histone gene regulation.* Genome Biol, 2006. **7**(12): p. R122.

127. Gabaldon, T. and E.V. Koonin, *Functional and evolutionary implications of gene orthology.* Nat Rev Genet, 2013. **14**(5): p. 360-6.

128. Koonin, E.V., *Orthologs, paralogs, and evolutionary genomics.* Annu Rev Genet, 2005. **39**: p. 309-38.

129. Richter, T.E., et al., *New rust resistance specificities associated with recombination in the Rp1 complex in maize.* Genetics, 1995. **141**(1): p. 373-81.

130. Leister, D., et al., *Rapid reorganization of resistance gene homologues in cereal genomes.* Proc Natl Acad Sci U S A, 1998. **95**(1): p. 370-5.

131. Nagy, E.D. and J.L. Bennetzen, *Pathogen corruption and site-directed recombination at a plant disease resistance gene cluster.* Genome Research, 2008. **18**(12): p. 1918-1923.

132. Gowik, U. and P. Westhoff, *The path from C3 to C4 photosynthesis.* Plant Physiol, 2011. **155**(1): p. 56-63.

133. Natale, D.A., et al., *Using the COG database to improve gene recognition in complete genomes.* Genetica, 2000. **108**(1): p. 9-17.

134. Olson, M.V., *When less is more: Gene loss as an engine of evolutionary change.* American Journal of Human Genetics, 1999. **64**(1): p. 18-23.

135. Tatusov, R.L., E.V. Koonin, and D.J. Lipman, *A genomic perspec-*

tive on protein families. Science, 1997. **278**(5338): p. 631-7.

136. Lynch, M. and J.S. Conery, *The evolutionary demography of duplicate genes.* J Struct Funct Genomics, 2003. **3**(1-4): p. 35-44.

137. Szathmary, E., F. Jordan, and C. Pal, *Molecular biology and evolution - Can genes explain biological complexity?* Science, 2001. **292**(5520): p. 1315-1316.

138. Lespinet, O., et al., *The role of lineage-specific gene family expansion in the evolution of eukaryotes.* Genome Res, 2002. **12**(7): p. 1048-59.

139. Lutfalla, G., et al., *Comparative genomic analysis reveals independent expansion of a lineage-specific gene family in vertebrates: the class II cytokine receptors and their ligands in mammals and fish.* BMC Genomics, 2003. **4**(1): p. 29.

140. Ranson, H., et al., *Evolution of supergene families associated with insecticide resistance.* Science, 2002. **298**(5591): p. 179-81.

141. Sonnhammer, E.L. and R. Durbin, *Analysis of protein domain families in Caenorhabditis elegans.* Genomics, 1997. **46**(2): p. 200-16.

142. Finn, R.D., et al., *Pfam: the protein families database.* Nucleic Acids Res, 2014. **42**(Database issue): p. D222-30.

143. Galperin, M.Y. and E.V. Koonin, *Searching for drug targets in microbial genomes.* Curr Opin Biotechnol, 1999. **10**(6): p. 571-8.

144. Natale, D.A., et al., *Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs).* Genome Biology, 2000. **1**(5).

145. Tatusov, R.L., et al., *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Res, 2000. **28**(1): p. 33-6.

146. Tatusov, R.L., et al., *The COG database: new developments in phylogenetic classification of proteins from complete genomes.* Nucleic Acids Res, 2001. **29**(1): p. 22-8.

147. Altschul, S.F., et al., *Basic local alignment search tool.* J Mol Biol, 1990. **215**(3): p. 403-10.

148. Doolittle, R.F., *The multiplicity of domains in proteins.* Annu Rev Biochem, 1995. **64**: p. 287-314.

149. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.

150. Remm, M., C.E. Storm, and E.L. Sonnhammer, *Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.* J Mol

Biol, 2001. **314**(5): p. 1041-52.

151. Lee, Y., et al., *Cross-referencing eukaryotic genomes: TIGR Orthologous Gene Alignments (TOGA).* Genome Res, 2002. **12**(3): p. 493-502.

152. Alexeyenko, A., et al., *Automatic clustering of orthologs and inparalogs shared by multiple proteomes.* Bioinformatics, 2006. **22**(14): p. e9-15.

153. Quackenbush, J., et al., *The TIGR gene indices: reconstruction and representation of expressed gene sequences.* Nucleic Acids Res, 2000. **28**(1): p. 141-5.

154. Quackenbush, J., et al., *The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species.* Nucleic Acids Res, 2001. **29**(1): p. 159-64.

155. ., V.D.S., *Graph clustering by flow simulation.* 2000, University of Utrecht, The Netherlands.

156. Warren, W.C., et al., *Genome analysis of the platypus reveals unique signatures of evolution.* Nature, 2008. **453**(7192): p. 175-83.

157. Wang, X., et al., *The genome of the mesopolyploid crop species Brassica rapa.* Nat Genet, 2011. **43**(10): p. 1035-9.

158. Amselem, J., et al., *Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea.* PLoS Genet, 2011. **7**(8): p. e1002230.

159. Batley, J. and D. Edwards, *Genome sequence data: management, storage, and visualization.* Biotechniques, 2009. **46**(5): p. 333-4, 336.

160. Kodama, Y., et al., *The Sequence Read Archive: explosive growth of sequencing data.* Nucleic Acids Res, 2012. **40**(Database issue): p. D54-6.

161. Cochrane, G., et al., *Facing growth in the European Nucleotide Archive.* Nucleic Acids Res, 2013. **41**(Database issue): p. D30-5.

162. Brandon, M.C., D.C. Wallace, and P. Baldi, *Data structures and compression algorithms for genomic sequence data.* Bioinformatics, 2009. **25**(14): p. 1731-8.

163. Nalbantoglu, O.U., D.J. Russell, and K. Sayood, *Data Compression Concepts and Algorithms and their Applications to Bioinformatics.* Entropy (Basel), 2010. **12**(1): p. 34.

164. Cochrane, G., C.E. Cook, and E. Birney, *The future of DNA sequence archiving.* Gigascience, 2012. **1**(1): p. 2.

165. *Statistics: Assembled/annotated sequence growth at EMBL-Bank.* 2014; Available from: http://www.ebi.ac.uk/ena/about/statistics.

166. Nakamura, Y., et al., *The International Nucleotide Sequence*

Database Collaboration. Nucleic Acids Res, 2013. **41**(Database issue): p. D21-4.

167. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2014. **42**(Database issue): p. D32-7.

168. Benson, D.A., et al., *GenBank.* Nucleic Acids Res, 2012. **40**(Database issue): p. D48-53.

169. Tateno, Y., et al., *DNA Data Bank of Japan (DDBJ) for genome scale research in life science.* Nucleic Acids Res, 2002. **30**(1): p. 27-30.

170. Stoesser, G., et al., *The EMBL Nucleotide Sequence Database.* Nucleic Acids Res, 1997. **25**(1): p. 7-14.

171. Leinonen, R., et al., *The European Nucleotide Archive.* Nucleic Acids Res, 2011. **39**(Database issue): p. D28-31.

172. Wheeler, D.L., et al., *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2008. **36**(Database issue): p. D13-21.

173. Stoesser, G., et al., *The EMBL Nucleotide Sequence Database.* Nucleic Acids Res, 2002. **30**(1): p. 21-6.

174. Hsi-Yang Fritz, M., et al., *Efficient storage of high throughput DNA sequencing data using reference-based compression.* Genome Res, 2011. **21**(5): p. 734-40.

175. Stein, L.D., et al., *The generic genome browser: a building block for a model organism system database.* Genome Res, 2002. **12**(10): p. 1599-610.

176. Grant, D., P. Cregan, and R.C. Shoemaker, *Genome organization in dicots: genome duplication in Arabidopsis and synteny between soybean and Arabidopsis.* Proc Natl Acad Sci U S A, 2000. **97**(8): p. 4168-73.

177. Mungall, C.J., D.B. Emmert, and C. FlyBase, *A Chado case study: an ontology-based modular schema for representing genome-associated biological information.* Bioinformatics, 2007. **23**(13): p. i337-46.

178. Skinner, M.E., et al., *JBrowse: a next-generation genome browser.* Genome Res, 2009. **19**(9): p. 1630-8.

179. Lewis, S.E., et al., *Apollo: a sequence annotation editor.* Genome Biol, 2002. **3**(12): p. RESEARCH0082.

180. *Generic Model Organism Database (GMOD).* Available from: http://gmod.org.

181. Ashburner, M. and R. Drysdale, *FlyBase–the Drosophila genetic database.* Development, 1994. **120**(7): p. 2077-9.

182. St Pierre, S.E., et al., *FlyBase 102–advanced approaches to inter-*

rogating FlyBase. Nucleic Acids Res, 2014. **42**(Database issue): p. D780-8.

183. Stein, L., et al., *WormBase: network access to the genome and biology of Caenorhabditis elegans.* Nucleic Acids Res, 2001. **29**(1): p. 82-6.

184. Harris, T.W., et al., *WormBase 2014: new views of curated biology.* Nucleic Acids Res, 2014. **42**(Database issue): p. D789-93.

185. Karolchik, D., et al., *The UCSC Genome Browser database: 2014 update.* Nucleic Acids Res, 2014. **42**(Database issue): p. D764-70.

186. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

187. Green, R.E., et al., *A draft sequence of the Neandertal genome.* Science, 2010. **328**(5979): p. 710-22.

188. Rosenbloom, K.R., et al., *ENCODE data in the UCSC Genome Browser: year 5 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D56-63.

189. *Encyclopedia of DNA Elements ENCODE.* 2014]; Available from: https://genome.ucsc.edu/ENCODE/.

190. Huala, E., et al., *The Arabidopsis Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant.* Nucleic Acids Res, 2001. **29**(1): p. 102-5.

191. Lamesch, P., et al., *The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1202-10.

192. *The Arabidopsis Information Resource.* [cited 2014; Available from: http://www.arabidopsis.org/.

193. Schoof, H., et al., *MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome.* Nucleic Acids Res, 2002. **30**(1): p. 91-3.

194. Goff, S.A., et al., *The iPlant Collaborative: Cyberinfrastructure for Plant Biology.* Front Plant Sci, 2011. **2**: p. 34.

195. Kawahara, Y., et al., *Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data.* Rice (N Y), 2013. **6**(1): p. 4.

196. Bombarely, A., et al., *The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1149-55.

197. Kersey, P.J., et al., *Ensembl Genomes 2013: scaling up access to genome-wide data.* Nucleic Acids Res, 2014. **42**(Database issue): p. D546-

52.

198. Goodstein, D.M., et al., *Phytozome: a comparative platform for green plant genomics.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1178-86.

199. Gonzales, M.D., et al., *The Legume Information System (LIS): an integrated information resource for comparative legume biology.* Nucleic Acids Res, 2005. **33**(Database issue): p. D660-5.

200. Ware, D., et al., *Gramene: a resource for comparative grass genomics.* Nucleic Acids Res, 2002. **30**(1): p. 103-5.

201. Monaco, M.K., et al., *Gramene 2013: comparative plant genomics resources.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1193-9.

202. Steinbach, D., et al., *GnpIS: an information system to integrate genetic and genomic data from plants and fungi.* Database (Oxford), 2013. **2013**: p. bat058.

203. Spannagl, M., et al., *MIPSPlantsDB–plant database resource for integrative and comparative plant genome research.* Nucleic Acids Res, 2007. **35**(Database issue): p. D834-40.

204. Schoof, H., et al., *MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics.* Nucleic Acids Res, 2004. **32**(Database issue): p. D373-6.

205. Arnold, R., et al., *SIMAP–the similarity matrix of proteins.* Bioinformatics, 2005. **21 Suppl 2**: p. ii42-6.

206. Arnold, R., et al., *SIMAP–the database of all-against-all protein sequence similarities and annotations with new interfaces and increased coverage.* Nucleic Acids Res, 2014. **42**(Database issue): p. D279-84.

207. Wilkinson, M.D. and M. Links, *BioMOBY: an open source biological web services proposal.* Brief Bioinform, 2002. **3**(4): p. 331-41.

208. Dowell, R.D., et al., *The distributed annotation system.* BMC Bioinformatics, 2001. **2**: p. 7.

209. Wilkinson, M., et al., *BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case.* Plant Physiol, 2005. **138**(1): p. 5-17.

210. Kawas, E., M. Senger, and M.D. Wilkinson, *BioMoby extensions to the Taverna workflow management and enactment software.* BMC Bioinformatics, 2006. **7**: p. 523.

211. Wilkinson, M., *Gbrowse Moby: a Web-based browser for BioMoby Services.* Source Code Biol Med, 2006. **1**: p. 4.

212. Lange, M., et al., *The LAILAPS search engine: a feature model for*

relevance ranking in life science databases. J Integr Bioinform, 2010. **7**(3).

213. Jenkinson, A.M., et al., *Integrating biological data–the Distributed Annotation System.* BMC Bioinformatics, 2008. **9 Suppl 8**: p. S3.

214. *GMOD in the Cloud.* [cited 2014; Available from: http://www.gmod.org/wiki/Cloud.

215. Lee, E., et al., *Web Apollo: a web-based genomic annotation editing platform.* Genome Biol, 2013. **14**(8): p. R93.

216. Dongen, v., *A cluster algorithm for graphs*, in *Technical Report INS-R0010.* 2000, National Research Institute for Mathematics and Computer Science in the Netherlands: Amsterdam.

217. Ashburner, M., et al., *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.

218. Jones, P., et al., *InterProScan 5: genome-scale protein function classification.* Bioinformatics, 2014. **30**(9): p. 1236-40.

219. Beissbarth, T. and T.P. Speed, *GOstat: find statistically overrepresented Gene Ontologies within a group of genes.* Bioinformatics, 2004. **20**(9): p. 1464-1465.

220. McCarthy, F.M., et al., *AgBase: a unified resource for functional analysis in agriculture.* Nucleic Acids Res, 2007. **35**(Database issue): p. D599-603.

221. Felsenstein, J., *PHYLIP (Phylogeny Inference Package).* 2005: Department of Genome Sciences, University of Washington, Seattle. p. Distributed by the author.

222. Matsumoto, T., et al., *Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries.* Plant Physiol, 2011. **156**(1): p. 20-8.

223. Lu, T.T., et al., *Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq.* Genome Research, 2010. **20**(9): p. 1238-1249.

224. Luo, M.C., et al., *A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor.* Proc Natl Acad Sci U S A, 2013. **110**(19): p. 7940-5.

225. Spannagl, M., et al., *Analysing complex Triticeae genomes - concepts and strategies.* Plant Methods, 2013. **9**(1): p. 35.

226. Frank, E., et al., *Data mining in bioinformatics using Weka.* Bioinformatics, 2004. **20**(15): p. 2479-2481.

227. Wicker, T., et al., *Frequent Gene Movement and Pseudogene Evolution Is Common to the Large and Complex Genomes of Wheat, Barley,*

and Their Relatives. Plant Cell, 2011. **23**(5): p. 1706-1718.

228. *Postgresql Database System.* 2014; Available from: http://www.postgresql.org/.

229. *JBOSS Application server.* 2014; Available from: http://jbossas.jboss.org/.

230. *Postgresql JDBC website.* 2014; Available from: https://jdbc.postgresql.org/.

231. Wikipedia. *XML Wikipedia website.* 2014; Available from: http://en.wikipedia.org/wiki/XML.

232. *Oracle Java Server Pages Website.* 2014; Available from: http://www.oracle.com/technetwork/java/javaee/jsp/index.html.

233. *Oracle Java Server Faces Website.* 2014; Available from: http://www.oracle.com/technetwork/java/javaee/javaserverfaces-139869.html.

234. Wikipedia. *Cascading Style Sheets Wikipedia website.* 2014; Available from: http://en.wikipedia.org/wiki/Cascading_Style_Sheets.

235. Wikipedia. *FTP Wikipedia website.* 2014; Available from: http://en.wikipedia.org/wiki/File_Transfer_Protocol.

236. Mascher, M., et al., *Anchoring and ordering NGS contig assemblies by population sequencing (POPSEQ).* Plant J, 2013. **76**(4): p. 718-27.

237. Rohe, M., et al., *The race-specific elicitor, NIP1, from the barley pathogen, Rhynchosporium secalis, determines avirulence on host plants of the Rrs1 resistance genotype.* EMBO J, 1995. **14**(17): p. 4168-77.

238. Goswami, R.S., et al., *Genomic analysis of host-pathogen interaction between Fusarium graminearum and wheat during early stages of disease development.* Microbiology-Sgm, 2006. **152**: p. 1877-1890.

239. Buschges, R., et al., *The barley Mlo gene: a novel control element of plant pathogen resistance.* Cell, 1997. **88**(5): p. 695-705.

240. Shirasu, K., et al., *A novel class of eukaryotic zinc-binding proteins is required for disease resistance signaling in barley and development in C-elegans.* Cell, 1999. **99**(4): p. 355-366.

241. Jung, J., et al., *The barley ERF-type transcription factor HvRAF confers enhanced pathogen resistance and salt tolerance in Arabidopsis.* Planta, 2007. **225**(3): p. 575-588.

242. Munns, R., et al., *Wheat grain yield on saline soils is improved by an ancestral Na+ transporter gene.* Nature Biotechnology, 2012. **30**(4): p. 360-U173.

243. Faris, J.D., et al., *A unique wheat disease resistance-like gene gov-*

erns effector-triggered susceptibility to necrotrophic pathogens. Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**(30): p. 13544-13549.

244. Saintenac, C., et al., *Identification of Wheat Gene Sr35 That Confers Resistance to Ug99 Stem Rust Race Group.* Science, 2013. **341**(6147): p. 783-786.

245. Singrun, C., et al., *Identification of powdery mildew and leaf rust resistance genes in common wheat (Triticum aestivum L.). Wheat varieties from the Caucasus, Central and Inner Asia.* Genetic Resources and Crop Evolution, 2004. **51**(4): p. 355-370.

246. Emebiri, L.C., et al., *Improvements in malting barley grain yield by manipulation of genes influencing grain protein content.* Euphytica, 2007. **156**(1-2): p. 185-194.

247. Wang, G.W., et al., *Association of barley photoperiod and vernalization genes with QTLs for flowering time and agronomic traits in a BC2DH population and a set of wild barley introgression lines.* Theoretical and Applied Genetics, 2010. **120**(8): p. 1559-1574.

248. Wu, X.S., X.P. Chang, and R.L. Jing, *Genetic Insight into Yield-Associated Traits of Wheat Grown in Multiple Rain-Fed Environments.* Plos One, 2012. **7**(2).

249. Flintham, J.E., et al., *Optimizing wheat grain yield: Effects of Rht (gibberellin-insensitive) dwarfing genes.* Journal of Agricultural Science, 1997. **128**: p. 11-25.

250. Mochida, K., et al., *TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics.* Plant Physiol, 2009. **150**(3): p. 1135-46.

251. van der Biezen, E.A. and J.D.G. Jones, *The NB-ARC domain: A novel signalling motif shared by plant resistance gene products and regulators of cell death in animals.* Current Biology, 1998. **8**(7): p. R226-R227.

252. Shewry, P.R. and N.G. Halford, *Cereal seed storage proteins: structures, properties and role in grain utilization.* Journal of Experimental Botany, 2002. **53**(370): p. 947-958.

253. Payne, P.I., *Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality.* Annual Review of Plant Physiology, 1987. **38**(1): p. 141-153.

254. Shewry, P.R., J.A. Napier, and A.S. Tatham, *Seed storage proteins: structures and biosynthesis.* Plant Cell, 1995. **7**(7): p. 945-56.

255. Shavrukov, Y., P. Langridge, and M. Tester, *Salinity tolerance and*

sodium exclusion in genus Triticum. Breeding Science, 2009. **59**(5): p. 671-678.

256. Wang, S.W., et al., *Wheat-Aegilops chromosome addition lines showing high iron and zinc contents in grains.* Breeding Science, 2011. **61**(2): p. 189-195.

257. Henry, R.J., *The carbohydrates of barley grains—A review.* Journal of the Institute of Brewing, 1988. **94**(2): p. 71-78.

258. Englyst, H.N., V. Anderson, and J.H. Cummings, *Starch and non-starch polysaccharides in some cereal foods.* J Sci Food Agric, 1983. **34**(12): p. 1434-40.

259. Jacobs, A.K., et al., *An Arabidopsis callose synthase, GSL5, is required for wound and papillary callose formation.* Plant Cell, 2003. **15**(11): p. 2503-2513.

260. Wei, B., et al., *Genome-wide analysis of the MADS-box gene family in Brachypodium distachyon.* PLoS One, 2014. **9**(1): p. e84781.

261. Tian, C., et al., *Genome-wide analysis of the GRAS gene family in rice and Arabidopsis.* Plant Mol Biol, 2004. **54**(4): p. 519-32.

262. Nakano, T., et al., *Genome-wide analysis of the ERF gene family in Arabidopsis and rice.* Plant Physiol, 2006. **140**(2): p. 411-32.

263. Shiu, S.H. and A.B. Bleecker, *Expansion of the receptor-like kinase/Pelle gene family and receptor-like proteins in Arabidopsis.* Plant Physiol, 2003. **132**(2): p. 530-43.

264. Silvar, C., et al., *Towards positional isolation of three quantitative trait loci conferring resistance to powdery mildew in two Spanish barley landraces.* PLoS One, 2013. **8**(6): p. e67336.

265. Munoz-Amatriain, M., et al., *Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome.* Genome Biol, 2013. **14**(6): p. R58.

266. Kugler, K.G., et al., *Quantitative trait loci-dependent analysis of a gene co-expression network associated with Fusarium head blight resistance in bread wheat (Triticum aestivum L.).* BMC Genomics, 2013. **14**: p. 728.

267. Ozkan, H., A.A. Levy, and M. Feldman, *Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group.* Plant Cell, 2001. **13**(8): p. 1735-47.

268. Haudry, A., et al., *Grinding up wheat: A massive loss of nucleotide diversity since domestication.* Molecular Biology and Evolution, 2007. **24**(7): p. 1506-1517.

269. Akhunov, E.D., et al., *Comparative analysis of syntenic genes in*

grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. Plant Physiol, 2013. **161**(1): p. 252-65.

270. Schnable, J.C., N.M. Springer, and M. Freeling, *Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(10): p. 4069-4074.

271. Flagel, L.E. and J.F. Wendel, *Evolutionary rate variation, genomic dominance and duplicate gene expression evolution during allotetraploid cotton speciation.* New Phytol, 2010. **186**(1): p. 184-93.

272. Chaudhary, B., et al., *Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium).* Genetics, 2009. **182**(2): p. 503-17.

273. Lin, J.Y., et al., *Structural and Functional Divergence of a 1-Mb Duplicated Region in the Soybean (Glycine max) Genome and Comparison to an Orthologous Region from Phaseolus vulgaris.* Plant Cell, 2010. **22**(8): p. 2545-2561.

274. Swigonova, Z., et al., *Close split of sorghum and maize genome progenitors.* Genome Res, 2004. **14**(10A): p. 1916-23.

275. Ownbey, M., *Natural hybridization and amphiploidy in the genus Tragopogon.* American Journal of Botany, 1950: p. 487-499.

276. Wang, J., et al., *Genomewide nonadditive gene regulation in Arabidopsis allotetraploids.* Genetics, 2006. **172**(1): p. 507-17.

277. Vigeland, M.D., et al., *Evidence for adaptive evolution of low-temperature stress response genes in a Pooideae grass ancestor.* New Phytol, 2013. **199**(4): p. 1060-8.

278. Jia, J., et al., *Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation.* Nature, 2013. **496**(7443): p. 91-5.

279. Ling, H.Q., et al., *Draft genome of the wheat A-genome progenitor Triticum urartu.* Nature, 2013. **496**(7443): p. 87-90.

280. Martin, W. and F. Salamini, *A meeting at the gene - Biodiversity and natural history.* Embo Reports, 2000. **1**(3): p. 208-210.

281. Badr, A., et al., *On the origin and domestication history of barley (Hordeum vulgare).* Molecular Biology and Evolution, 2000. **17**(4): p. 499-510.

282. Heun, M., et al., *Site of einkorn wheat domestication identified by DNA fingerprinting.* Science, 1997. **278**(5341): p. 1312-1314.

283. Salamini, F., et al., *Genetics and geography of wild cereal domesti-*

cation in the Near East. Nature Reviews Genetics, 2002. **3**(6): p. 429-441.

284. Peng, J.H.H., D.F. Sun, and E. Nevo, *Domestication evolution, genetics and genomics in wheat.* Molecular Breeding, 2011. **28**(3): p. 281-301.

285. Shewry, P., N. Halford, and A. Tatham, *High molecular weight subunits of wheat glutenin.* Journal of Cereal Science, 1992. **15**(2): p. 105-120.

286. Dong, Z., et al., *Haplotype variation of Glu-D1 locus and the origin of Glu-D1d allele conferring superior end-use qualities in common wheat.* PLoS One, 2013. **8**(9): p. e74859.

287. Simons, K., et al., *Genetic Mapping Analysis of Bread-Making Quality Traits in Spring Wheat.* Crop Science, 2012. **52**(5): p. 2182-2197.

288. Yahiaoui, N., et al., *Genome analysis at different ploidy levels allows cloning of the powdery mildew resistance gene Pm3b from hexaploid wheat.* Plant Journal, 2004. **37**(4): p. 528-538.

289. Cuthbert, P.A., D.J. Somers, and A. Brule-Babel, *Mapping of Fhb2 on chromosome 6BS: a gene controlling Fusarium head blight field resistance in bread wheat (Triticum aestivum L.).* Theoretical and Applied Genetics, 2007. **114**(3): p. 429-437.

290. Gupta, P.K., et al., *Wheat genomics: present status and future prospects.* Int J Plant Genomics, 2008. **2008**: p. 896451.

291. Schreiber, A.W., et al., *Transcriptome-scale homoeolog-specific transcript assemblies of bread wheat.* BMC Genomics, 2012. **13**: p. 492.

292. Chevreux, B., et al., *Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs.* Genome Research, 2004. **14**(6): p. 1147-1159.

293. Zerbino, D.R., et al., *Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler.* PLoS One, 2009. **4**(12): p. e8407.

294. Bansal, V. and V. Bafna, *HapCUT: an efficient and accurate algorithm for the haplotype assembly problem.* Bioinformatics, 2008. **24**(16): p. i153-9.

295. Dolezel, J., et al., *Chromosomes in the flow to simplify genome analysis.* Funct Integr Genomics, 2012. **12**(3): p. 397-416.

296. Vrana, J., et al., *Flow cytometric chromosome sorting in plants: the next generation.* Methods, 2012. **57**(3): p. 331-7.

297. Ma, X.F. and J.P. Gustafson, *Genome evolution of allopolyploids: a process of cytological and genetic diploidization.* Cytogenet Genome Res,

2005. **109**(1-3): p. 236-49.

298. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.

299. Kuroshu, R.M., et al., *Cost-effective sequencing of full-length cDNA clones powered by a de novo-reference hybrid assembly.* PLoS One, 2010. **5**(5): p. e10517.

300. Hatey, F., et al., *Expressed sequence tags for genes: a review.* Genetics Selection Evolution, 1998. **30**(6): p. 521-541.

301. Ozsolak, F. and P.M. Milos, *RNA sequencing: advances, challenges and opportunities.* Nat Rev Genet, 2011. **12**(2): p. 87-98.

302. *trans-National Infrastructure for Plant Genomic Science.* [cited 2014; Available from: http://www.transplantdb.eu/.

303. *Semantic Web.* [cited 2014; Available from: http://semanticweb.org.

304. Day, D.A. and M.F. Tuite, *Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview.* Journal of Endocrinology, 1998. **157**(3): p. 361-371.

305. Kaufmann, K., A. Pajoro, and G.C. Angenent, *Regulation of transcription in plants: mechanisms controlling developmental switches.* Nature Reviews Genetics, 2010. **11**(12): p. 830-842.

306. Coulon, A., et al., *Eukaryotic transcriptional dynamics: from single molecules to cell populations.* Nature Reviews Genetics, 2013. **14**(8): p. 572-584.

307. GuhaThakurta, D., *Computational identification of transcriptional regulatory elements in DNA sequence.* Nucleic Acids Res, 2006. **34**(12): p. 3585-98.

308. Loots, G.G., *Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis.* Adv Genet, 2008. **61**: p. 269-93.

309. Cooper, S.J., et al., *Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome.* Genome Res, 2006. **16**(1): p. 1-10.

310. Kumari, S. and D. Ware, *Genome-wide computational prediction and analysis of core promoter elements across plant monocots and dicots.* PLoS One, 2013. **8**(10): p. e79011.

311. Wang, Y., T. Hindemitt, and K.F. Mayer, *Significant sequence similarities in promoters and precursors of Arabidopsis thaliana non-conserved*

microRNAs. Bioinformatics, 2006. **22**(21): p. 2585-9.

312. Arkhipova, I.R., *Promoter elements in Drosophila melanogaster revealed by sequence analysis.* Genetics, 1995. **139**(3): p. 1359-69.

313. Wasserman, W.W. and A. Sandelin, *Applied bioinformatics for the identification of regulatory elements.* Nat Rev Genet, 2004. **5**(4): p. 276-87.

314. Ohler, U. and H. Niemann, *Identification and analysis of eukaryotic promoters: recent computational approaches.* Trends Genet, 2001. **17**(2): p. 56-60.

315. Blanchette, M. and M. Tompa, *Discovery of regulatory elements by a computational method for phylogenetic footprinting.* Genome Res, 2002. **12**(5): p. 739-48.

316. Ganley, A.R. and T. Kobayashi, *Phylogenetic footprinting to find functional DNA elements.* Methods Mol Biol, 2007. **395**: p. 367-80.

317. Blanchette, M. and M. Tompa, *FootPrinter: A program designed for phylogenetic footprinting.* Nucleic Acids Res, 2003. **31**(13): p. 3840-2.

318. Sandelin, A., W.W. Wasserman, and B. Lenhard, *ConSite: web-based prediction of regulatory elements using cross-species comparison.* Nucleic Acids Research, 2004. **32**: p. W249-W252.

319. van Nimwegen, E., *Finding regulatory elements and regulatory motifs: a general probabilistic framework.* BMC Bioinformatics, 2007. **8 Suppl 6**: p. S4.

320. Higo, K., et al., *Plant cis-acting regulatory DNA elements (PLACE) database: 1999.* Nucleic Acids Res, 1999. **27**(1): p. 297-300.

321. Liu, X., D.L. Brutlag, and J.S. Liu, *BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.* Pac Symp Biocomput, 2001: p. 127-38.

322. Kim, N.K., K. Tharakaraman, and J.L. Spouge, *Adding sequence context to a Markov background model improves the identification of regulatory elements.* Bioinformatics, 2006. **22**(23): p. 2870-5.

323. Elemento, O., N. Slonim, and S. Tavazoie, *A universal framework for regulatory element discovery across all Genomes and data types.* Molecular Cell, 2007. **28**(2): p. 337-350.

324. Young, J.A., et al., *In silico discovery of transcription regulatory elements in Plasmodium falciparum.* BMC Genomics, 2008. **9**: p. 70.

325. He, L. and G.J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation.* Nat Rev Genet, 2004. **5**(7): p. 522-31.

326. Brennecke, J., et al., *bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene*

hid in Drosophila. Cell, 2003. **113**(1): p. 25-36.

327. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.

328. Bentwich, I., et al., *Identification of hundreds of conserved and nonconserved human microRNAs.* Nat Genet, 2005. **37**(7): p. 766-70.

329. Jones-Rhoades, M.W., D.P. Bartel, and B. Bartel, *MicroRNAs and their regulatory roles in plants.* Annual Review of Plant Biology, 2006. **57**: p. 19-53.

330. Eckardt, N.A., *A microRNA cascade in plant defense.* Plant Cell, 2012. **24**(3): p. 840.

331. Kozomara, A. and S. Griffiths-Jones, *miRBase: integrating microRNA annotation and deep-sequencing data.* Nucleic Acids Res, 2011. **39**(Database issue): p. D152-7.

332. Grennan, A.K., *Arabidopsis MicroRNAs.* Plant Physiol, 2008. **146**(1): p. 3-4.

333. Reinhart, B.J., et al., *MicroRNAs in plants.* Genes Dev, 2002. **16**(13): p. 1616-26.

334. Zheng, H., et al., *Advances in the Techniques for the Prediction of microRNA Targets.* Int J Mol Sci, 2013. **14**(4): p. 8179-87.

335. Tempel, S. and F. Tahi, *A fast ab-initio method for predicting miRNA precursors in genomes.* Nucleic Acids Res, 2012. **40**(11): p. e80.

336. Lim, L.P., et al., *Vertebrate MicroRNA genes.* Science, 2003. **299**(5612): p. 1540-1540.

337. Zhang, L., et al., *A genome-wide characterization of microRNA genes in maize.* PLoS Genet, 2009. **5**(11): p. e1000716.

338. Bell, O., et al., *Determinants and dynamics of genome accessibility.* Nat Rev Genet, 2011. **12**(8): p. 554-64.

339. Zentner, G.E. and S. Henikoff, *Regulation of nucleosome dynamics by histone modifications.* Nat Struct Mol Biol, 2013. **20**(3): p. 259-66.

340. Cedar, H. and Y. Bergman, *Linking DNA methylation and histone modification: patterns and paradigms.* Nat Rev Genet, 2009. **10**(5): p. 295-304.

341. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nature Genetics, 2003. **33**: p. 245-254.

342. Shen, H.S., et al., *Genome-Wide Analysis of DNA Methylation and Gene Expression Changes in Two Arabidopsis Ecotypes and Their Reciprocal*

Hybrids. Plant Cell, 2012. **24**(3): p. 875-892.

343. Bird, A., *DNA methylation patterns and epigenetic memory.* Genes Dev, 2002. **16**(1): p. 6-21.

344. Martienssen, R.A. and V. Colot, *DNA methylation and epigenetic inheritance in plants and filamentous fungi.* Science, 2001. **293**(5532): p. 1070-4.

345. Chan, S.W., I.R. Henderson, and S.E. Jacobsen, *Gardening the genome: DNA methylation in Arabidopsis thaliana.* Nat Rev Genet, 2005. **6**(5): p. 351-60.

346. Law, J.A. and S.E. Jacobsen, *Establishing, maintaining and modifying DNA methylation patterns in plants and animals.* Nat Rev Genet, 2010. **11**(3): p. 204-20.

347. Zilberman, D., et al., *Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription.* Nature Genetics, 2007. **39**(1): p. 61-69.

348. Jayatilake, D.V., et al., *Genetic mapping and marker development for resistance of wheat against the root lesion nematode Pratylenchus neglectus.* BMC Plant Biol, 2013. **13**: p. 230.

349. McCarthy, M.I., et al., *Genome-wide association studies for complex traits: consensus, uncertainty and challenges.* Nat Rev Genet, 2008. **9**(5): p. 356-69.

350. McCarthy, M.I. and J.N. Hirschhorn, *Genome-wide association studies: past, present and future.* Hum Mol Genet, 2008. **17**(R2): p. R100-1.

351. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.

352. Weigel, D. and R. Mott, *The 1001 genomes project for Arabidopsis thaliana.* Genome Biol, 2009. **10**(5): p. 107.

353. Hardy, J. and A. Singleton, *Genomewide association studies and human disease.* N Engl J Med, 2009. **360**(17): p. 1759-68.

354. Wellcome Trust Case Control, C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.

355. Huang, X., et al., *Genome-wide association studies of 14 agronomic traits in rice landraces.* Nat Genet, 2010. **42**(11): p. 961-7.

356. Li, Y., et al., *Association mapping of local climate-sensitive quantitative trait loci in Arabidopsis thaliana.* Proc Natl Acad Sci U S A, 2010.

107(49): p. 21199-204.

357. Pasam, R.K., et al., *Genome-wide association studies for agronomical traits in a world wide spring barley collection.* Bmc Plant Biology, 2012. **12**.

358. Neumann, K., et al., *Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.).* Molecular Breeding, 2011. **27**(1): p. 37-58.

359. Brachi, B., G.P. Morris, and J.O. Borevitz, *Genome-wide association studies in plants: the missing heritability is in the field.* Genome Biology, 2011. **12**(10).

360. Nordborg, M. and D. Weigel, *Next-generation genetics in plants.* Nature, 2008. **456**(7223): p. 720-3.

361. Zhu, C.S., et al., *Status and Prospects of Association Mapping in Plants.* Plant Genome, 2008. **1**(1): p. 5-20.

362. Tian, F., et al., *Genome-wide association study of leaf architecture in the maize nested association mapping population.* Nat Genet, 2011. **43**(2): p. 159-62.

363. Huang, X., T. Lu, and B. Han, *Resequencing rice genomes: an emerging new era of rice genomics.* Trends Genet, 2013. **29**(4): p. 225-32.

364. Xu, X., et al., *Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes.* Nat Biotechnol, 2012. **30**(1): p. 105-11.

365. Huang, X. and B. Han, *A crop of maize variants.* Nat Genet, 2012. **44**(7): p. 734-5.

366. Gore, M.A., et al., *A first-generation haplotype map of maize.* Science, 2009. **326**(5956): p. 1115-7.

367. Chia, J.M., et al., *Maize HapMap2 identifies extant variation from a genome in flux.* Nat Genet, 2012. **44**(7): p. 803-7.

368. Hufford, M.B., et al., *Comparative population genomics of maize domestication and improvement.* Nat Genet, 2012. **44**(7): p. 808-11.

369. Yang, W.P., et al., *Comparison of DNA marker technologies in characterizing plant genome diversity: Variability in Chinese sorghums.* Crop Science, 1996. **36**(6): p. 1669-1676.

370. Lee, D., J.C. Reeves, and R.J. Cooke, *DNA profiling and plant variety registration .1. The use of random amplified DNA polymorphisms to discriminate between varieties of oilseed rape.* Electrophoresis, 1996. **17**(1): p. 261-265.

371. Powell, W., G.C. Machray, and J. Provan, *Polymorphism revealed*

by simple sequence repeats. Trends in Plant Science, 1996. **1**(7): p. 215-222.

372. Shapiro, J.A. and R. von Sternberg, *Why repetitive DNA is essential to genome function.* Biological Reviews, 2005. **80**(2): p. 227-250.

373. Marra, M., et al., *zA map for sequence analysis of the Arabidopsis thaliana genome.* Nat Genet, 1999. **22**(3): p. 265-70.

374. Schulte, D., et al., *BAC library resources for map-based cloning and physical map construction in barley (Hordeum vulgare L.).* BMC Genomics, 2011. **12**: p. 247.

375. Schadt, E.E., S. Turner, and A. Kasarskis, *A window into third-generation sequencing.* Human Molecular Genetics, 2010. **19**: p. R227-R240.

376. Eid, J., et al., *Real-Time DNA Sequencing from Single Polymerase Molecules.* Science, 2009. **323**(5910): p. 133-138.

377. Hayden, E.C., *Nanopore genome sequencer makes its debut.* Nature, 2012.

378. Li, F., et al., *Genome sequence of the cultivated cotton Gossypium arboreum.* Nat Genet, 2014. **46**(6): p. 567-72.

379. Choulet, F., et al., *Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces.* Plant Cell, 2010. **22**(6): p. 1686-701.

380. Duan, J., et al., *Optimizing de novo common wheat transcriptome assembly using short-read RNA-Seq data.* BMC Genomics, 2012. **13**: p. 392.

381. Li, C. and J. Dubcovsky, *Wheat FT protein regulates VRN1 transcription through interactions with FDL2.* Plant J, 2008. **55**(4): p. 543-54.

382. Lee, B. and D. Lee, *Protein comparison at the domain architecture level.* Bmc Bioinformatics, 2009. **10**.

383. Song, N., et al., *Sequence similarity network reveals common ancestry of multidomain proteins.* Plos Computational Biology, 2008. **4**(5).