HelmholtzZentrum münchen
Deutsches Forschungszentrum für Gesundheit und Umwelt

PGSB Plant Genome and Systems Biology

Technische Universität München

The genome and transcriptome of Triticeae genomes and the impact of polyploidization

Matthias Franz Xaver Pfeifer

2014

Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt

Lehrstuhl für Genomorientierte Bioinformatik

The genome and transcriptome of Triticeae genomes and the impact of polyploidization

Matthias Franz Xaver Pfeifer

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. Claus Schwechheimer

Prüfer der Dissertation:

1. Univ.-Prof. Dr. Hans-Werner Mewes
2. Univ.-Prof. Dr. Chris-Carolin Schön
3. Prof. Michael Bevan, Ph.D.
   University of East Anglia, Norwich, Norfolk, UK

Die Dissertation wurde am 31. Juli 2014 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 10. Dezember 2014 angenommen.

Danksagung

Diese Doktorarbeit war nur durch die große Unterstützung vieler Personen möglich, an die ich sehr gerne ein paar persönliche Worte des Dankes richten möchte:

Zunächst möchte ich mich bei Prof. Dr. Hans-Werner Mewes bedanken, der meine Promotion an der TU München und im Rahmen meiner Tätigkeit am Helmholtz Zentrum München begleitet hat. Bei Prof. Dr. Claus Schwechheimer für die Übernahme des Prüfungsvorsitzes und bei Prof. Dr. Chris-Carolin Schön und Prof. Dr. Michael Bevan für die Begutachtung dieser Arbeit.

Ein herzliches „vergelt's Gott" vor allem Dr. Klaus Mayer, der mir seit mehr als drei Jahren die spannende und abwechslungsreiche Arbeit mit „grünen" Genomen und Transkriptomen in zahlreichen Projekten ermöglicht. Vielen Dank, dass du mich bei meiner Doktorarbeit immer unterstützt hast und mir ein „wissenschaftlicher Lehrmeister" und wertvoller Ansprechpartner warst.

Ich möchte mich auch bei allen meinen Kollegen für zahlreiche interessante und hilfreiche Diskussionen und eine motivierende und gleichzeitig sehr angenehme Arbeitsatmosphäre bedanken. Dabei im Besonderen bei Mihaela, die mich auf meinen ersten Schritten in der Welt der Pflanzengenome begleitet hat, bei Manuel, für seine Unterstützung bei der Genfamilienanalyse und der funktionellen Genannotation, und bei Karl, der mich in die Analyse biologischer Netzwerke eingeführt hat, auf alle meine statistischen Fragen eine Antwort wusste und mit dem ich gemeinsam viele Stunden über die Geheimnisse der Weizenkornentwicklung gegrübelt habe.

Ein großer Dank gilt allen externen Kollaborationspartnern, den Forschungsgruppen unter Leitung von Prof. Dr. Neil Hall, Prof. Dr. Michael Bevan und Prof. Dr. Odd-Arne Olsen, sowie dem Internationalen Weizen-Genom-Sequenzierungskonsortium.

Ein besonderes Dankeschön gilt meiner gesamten Familie und meinen Freunden, die meine Arbeit immer mit großem Interesse verfolgt haben. Mama und Papa, vielen Dank, dass ihr mich ausnahmslos auf jeder Etappe meiner schulischen, beruflichen und akademischen Ausbildung begleitet habt. Ihr seit mir immer mit Rat und Tat zur Seite gestanden und habt dadurch ermöglicht, dass ich meine Ziele erfolgreich verwirklichen konnte.

Liebe Michaela, deine uneingeschränkte Geduld und deine bedingungslose Unterstützung waren ein großer Rückhalt während meiner Promotion. Vor allem aber auch vielen Dank für alle gemeinsamen Momente, die mich meine Arbeit und den ein oder anderen langen Arbeitstag vergessen ließen. Ich bin unendlich glücklich dich an meiner Seite zu wissen!

Abstract

Tremendous population growth, serious impacts of climate change and globally endangered ecosystems require sustainable intensification of agricultural productivity. Therefore, international efforts have been dedicated to enhance genome-driven research strategies to accelerate improvement of crop varieties. So far, large genome sizes, high sequence repetitivity and complex genome constitutions have delayed the comprehensive analyses of Triticeae genomes including the major cereals rye, barley and wheat. Especially the allohexaploid genome structure of bread wheat (*Triticum aestivum*) impeded the development of genome-wide resources needed for a function- and systems-level understanding of the genome biology and transcriptional regulatory mechanisms of one of the world's most important crop. By exploiting advances in high-throughput genome and transcriptome sequencing technologies, this thesis presents bioinformatic strategies for overcoming those barriers and for generating genomic resources for the complete gene catalogue of hexaploid wheat.

To assemble the protein-coding space of the 17 gigabase pair large bread wheat genome, while maintaining highly similar homoeologous sequences as distinct copies, I developed and implemented a novel computational strategy integrating whole genome shotgun sequences into an orthologous gene family framework. Comparative analyses of the gene repertoire of bread wheat with the orthologous gene family sizes in the reference grasses *Brachypodium distachyon*, *Oryza sativa* and *Sorghum bicolor* as well as in *Aegilops tauschii*, the diploid wheat D-genome progenitor, revealed considerable genome dynamics including an abundance of pseudogenes and gene fragments. Despite a substantial retention of homoeologous genes in single-copy gene families and a general reduction of large gene family sizes in hexaploid wheat, various expanded gene families were identified after polyploidization. This expansion might be a result of domestication as the affected gene families were related to agriculturally important traits and crop productivity including defence response and disease resistance, energy metabolism and photosynthesis as well as compounds controlling grain filling and maturation.

Chromosome flow cytometry allowed the International Wheat Genome Sequence Consortium to isolate, sequence and assemble DNA of individual wheat chromosome arms, thus facilitated the homoeologous-specific annotation of gene sequences and structures. By incorporating extrinsic sequence information from closely related grass genomes and wheat transcriptome data sets, an overall comparable number of protein-coding genes were identified across the homoeologous genomes in this thesis. However, differences were observed in the gene density, the syntenic conservation and the gene family constitutions for individual chromosomes and chromosome arms. In addition to 124,201 high-confidence genes with homology support from other plant genomes, thousands of deteriorated, potentially non-functional gene ruins were detected, which indicated high activity of pseudogenization mechanisms in Triticeae genomes. Deep RNA-sequencing revealed abundance of novel (non-protein-coding) transcriptional active regions and extensive tissue-specific alternative splicing, frequently generating premature termination codon-containing transcripts. Conservation of the observed splicing patterns across the wheat genomes suggested that post-transcriptional processing constitutes an additional important regulatory layer in Triticeae genomes.

By using these previously unknown genomic resources I investigated the gene expression in different cell types of developing endosperm and elucidated the contributions of homoeologous transcripts to the entire wheat grain transcriptome. Partitioning of gene expression for homoeologs in the spatiotemporal progression of grain development indicated subfunctionalization of redundant gene copies or pre-existing transcriptional differences in the parental genomes. Rather than global transcriptional dominance, the observed gene expression differences were dependent on cell type and developmental stage. Functional compartmentalization of the wheat transcriptome and genome asymmetry for single gene families affecting baking quality suggested that individual genomes contribute differently to specific cellular functions and agricultural important traits. The organization of the wheat genome into transcriptional active chromosomal domains, often associated with homoeologous gene expression bias and genome dominance, indicated a complex regulatory interplay of genetic and epigentic mechanisms orchestrating gene expression in a polyploid cereal.

This thesis provides novel insights in the genome architecture and transcriptional organization of bread wheat, the agricultural most important Triticeae. The established genomic resources will support to gain a better understanding of the biological mechanisms that control a polyploid cereal genome and will enable both system-level and targeted analyses of single genes or gene families and their association to traits of economic and scientific interest.

Zusammenfassung

Weltweites Bevölkerungswachstum und ein sich änderndes Klima verlangen eine nachhaltige Anpassung aktueller landwirtschaftlicher Produktionsweisen. Diese sollen zu einer deutlichen Ertragssteigerung führen, jedoch gleichzeitig bedrohte Ökosysteme schützen und einen Verlust an Biodiversität vermeiden. Vor allem genomorientierte Forschungsansätze sind ein wesentlicher Bestandteil für eine beschleunigte Verbesserung bestehender Getreidesorten. Bislang wurden diese Strategien jedoch durch einen hohen Anteil repetitiver Sequenzelemente und einer komplexen Genomstruktur der Triticeae, zu denen Gerste, Roggen und Weizen zählen, erschwert. Besonders die allohexaploide genetische Ausstattung des Brotweizens (*Triticum aestivum*) hat die Entwicklung genomischer Ressourcen verlangsamt und ein umfassendes Verständnis über die Genomik und Systembiologie eines der weltweit wichtigsten Getreidearten beeinträchtigt. Diese Arbeit stellt neue bioinformatische Ansätze für die Erstellung eines umfassenden Genkataloges des Brotweizengenoms vor. Um die bisherigen Schwierigkeiten in der Analyse des hexaploiden Brotweizengenoms zu überwinden, wurden unterschiedliche Hochdurchsatzsequenzierungsdaten verwendet.

Um den proteinkodierenden Anteil der 17 Milliarden DNA-Bausteine des Brotweizengenoms zu charakterisieren wurde in dieser Arbeit ein neuartiges Assemblierungsverfahren implementiert. Dieses integriert genomische Sequenzfragmente aus *Whole-Genome-Shutgun*-Sequenzierungen in ein orthologes Genfamilien-Gerüst, um zusammengehörende genomische Sequenzfragmente zu assemblieren und gleichzeitig homoeologe Kopien, die zueinander eine hohe Sequenzähnlichkeit aufweisen, zu unterscheiden. Vergleichende Analysen des Genkataloges von Brotweizen mit der Genfamilienzusammensetzung in *Brachypodium distachyon*, *Oryza sativa* und *Sorghum bicolor*, sowie mit *Aegilops tauschii*, dem diploiden Vorgänger des Weizen-D-Genoms, deuten auf ein hohes Maß an Genomplastizität hin, geprägt durch eine Vielzahl von Pseudogenen und Genfragmenten. Nach der Polyploidisierung wurden polyploide Gene in kleinen Genfamilien meist erhalten, wohingegen für größere Genfamilien tendenziell eine Reduktion der Genanzahl festgestellt wurde. Desweiteren konnten zahlreiche Genfamilien ermittelt werden, die eine deutlich erhöhte Anzahl an Genkopien im hexaploiden Weizen aufweisen. Diese konnten mit wichtigen Getreideeigenschaften wie Widerstandsfähigkeit, Energiestoffwechsel, Photosynthese und Kornentwicklung in Verbindung gebracht werden und reflektieren somit möglicherweise Selektionseffekte während der Züchtung und Kultivierung des Brotweizens.

Durchflusszytometrie ermöglichte eine getrennte Sequenzierung und Assemblierung einzelner Brotweizen-Chromosomenarmen und somit die genomspezifische Annotation von Genstrukturen und –sequenzen im Rahmen des Internationalen Weizen-Genom-Sequenzierungskonsortium. Unter Berücksichtigung von Proteinsequenzen nahverwandter Gräser und Weizen Transkriptom-Sequenzierungsdaten konnte in dieser Arbeit eine gesamtheitlich ausgeglichene genetische Ausstattung der Genome bestimmt werden. Allerdings wurden auch deutliche Unterschiede in der Gendichte, der Syntenie und der Zusammensetzung von Genfamilien zwischen einzelnen Chromosomenarmen und Chromosomen festgestellt. Zusätzlich zu 124201 proteinkodierenden Genen wurde eine große Zahl an degenerierten Genfragmenten gefunden. Dies lässt auf eine hohe Aktivität Pseudogen-erzeugender Mechanismen in Triticeae-Genomen schließen. Eine Vielzahl von neuartigen (nicht-proteinkodierenden) transkribierten Sequenzbereichen wurden mittels Hochdurchsatz-Transkriptom-Sequenzierung definiert. Außerdem wurde ein hohes Maß an alternativen Spleißen gefunden, welches häufig in Transkripten mit vorzeitigem Translationsende resultierte und in großem Umfang für homoeologe Genkopien konserviert war. Diese Ergebnisse deuten auf mögliche regulatorische Funktionen der posttranskriptionalen Modifikation in Triticeae-Genomen hin.

Mittels des erstellten Sequenzentwurfs einzelner Brotweizen-Chromosomen und der darauf basierenden Genannotation untersucht diese Arbeit im Weiteren die Genexpression in verschiedenen Zelltypen und Stadien der Weizenkornentwicklung. Dabei wurde ein besonderer Fokus auf die Regulation, das Verhalten und den Beitrag homoeologer Genkopien gelegt. Für durch Polyploidisierung duplizierten Gene wurde eine spezifische Aktivität zu unterschiedlichen Entwicklungsstadien gefunden, was auf eine teilweise Subfunktionalisierung oder bereits bestehende Unterschiede in den Elterngenomen schließen lässt. Es konnten keine Anzeichen für eine genomweite Dominanz eines Genoms festgestellt werden, wohingegen zelltyp- und zeitpunktbestimmte Expressionsunterschiede zwischen homoeologen Genen deutlich wurden. Eine asymmetrische transkriptionelle Regulation einzelner Genome konnte für einzelne molekulare Funktionen im gesamtheitlichen Weizentranskriptom, sowie in einer gezielten Analyse von Genfamilien, die zu den charakteristischen Merkmalen und Backeigenschaften des Brotweizens beitragen, ausgemacht werden. Diese Beobachtungen deuten auf eine Aufgabenverteilung in der Weizenkornentwicklung hin und lassen somit spezifische landwirtschaftlich wichtige molekulare Eigenschaften einzelner Genome und Genkopien zuordnen. Genomspezifische transkriptionelle Unterschiede, welche oftmals auch für chromosomale Domänen gefunden wurden, lassen auf ein komplexes regulatorisches Wechselspiel von genetischen und epigenetischen Mechanismen zur Steuerung der Genexpression im polyploiden Brotweizen schließen.

Diese Arbeit gibt neue Erkenntnisse in die genetische Ausstattung und transkriptionelle Organisation des großen polyploiden Genoms von Brotweizen, dem landwirtschaftlichen und industriell wichtigsten Vertreter der Triticeae. Die vorgestellten Werkzeuge und genomischen Ressourcen bilden die Basis für weitere globale und systembiologische Analysen sowie die gezielte Analyse einzelner Genfamilien und Gene, wodurch diese Arbeit zu einem detaillierteren Verständnis der biologischen Mechanismen in einem wichtigen polyploiden Getreidegenom beiträgt.

Scientific publications

The following overview summarizes all peer-reviewed publications of projects in which I was involved in during my dissertation. The major scientific publications related to the presented results are briefly introduced and personal contributions to the individual studies highlighted.

Joint first authorships are marked with ‡.

2012

A transcriptome map of perennial ryegrass (*Lolium perenne* L.)

B. Studer, S. Byrne, R. O. Nielsen, F. Panitz, C. Bendixen, M. S. Islam, **M. Pfeifer**, T. Lübberstedt and T. Asp

BMC Genomics. 13(1):140, 2012.

Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content

P. Hernandez, M. M. Martis, G. Dorado, **M. Pfeifer**, S. Gàlvez, S. Schaaf, N. Jouve, H. Simkovà, M. Valàrik, J. Dolezel and K. F. X. Mayer

The Plant Journal. 69:377–386, 2012.

A series of inter- and intra-chromosomal rearrangements shaped the structures of chromosomes 4A, 5A and 7B of bread wheat (*Triticum aestivum*). On basis of second generation sequencing reads generated for flow-sorted chromosome arms and by exploiting conserved synteny between bread wheat and *Brachypodium*, rice and sorghum, this paper reports the bioinformatic analysis of the evolutionary history for these chromosomes.

Personal contributions: In this work I contributed to the computational analysis of the wheat gene content of chromosome 4A. The findings in this study corroborate the observations reported in **Chapter 4** of this thesis, which will provide further insights in the homoeologous relationships for these chromosomes.

Analysis of the bread wheat genome using whole-genome shotgun sequencing

R. Brenchley‡, M. Spannagl‡, **M. Pfeifer‡**, G. L. A. Barker‡, R. D'Amore‡, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou, D. Bolser, S. Kay, D. Waite, M. Trick, I. Bancroft, Y. Gu, N. Huo, M. C. Luo, S. Sehgal, B. Gill, S. Kianian, O. Anderson, P. Kersey, J. Dvorak, W. R. McCombie, A. Hall, K. F. X. Mayer, K. J. Edwards, M. W. Bevan and N. Hall

Nature. 491(7426):705–710, 2012.

High sequence similarity among homoeologous gene copies represents a major barrier for the analysis of the bread wheat genome (*Triticum aestivum*). This manuscript presents a novel comparative genomics-based assembly protocol, which enabled distinguishing homoeologous genomic shotgun sequences and quantifying gene family sizes and gene copy numbers for bread wheat. Comparative analysis between the gene family sizes of hexaploid wheat and diploid reference grasses as well as the diploid D-genome progenitor *Aegilops tauschii* revealed a dynamic genome content with retention of homoeologous single-copy genes, a general trend for gene loss in larger gene families, presence of gene families with expanded copy numbers related to agricultural important traits and an abundance of gene fragments.

Personal contributions: This study constitutes the major publication for **Chapter 3** of my thesis. On basis of an orthologous gene family framework, which was defined by known protein sequences of related grass genomes, I designed, implemented and performed the orthologous group assembly by using whole genome sequencing reads obtained for the genomes of hexaploid bread wheat and diploid *Aegilops tauschii*. Moreover, I conducted two *in silico* experiments to calibrate the assembly parameters and to evaluate the presented procedure. Furthermore, I was responsible for the computation and statistical analysis of the wheat gene copy number and the identification and analysis of sub-assemblies forming ”stacks”. Additionally, I contributed to the genome-of-origin assignment for the generated wheat sub-assemblies by providing the training data set utilized in the machine learning algorithm.

A physical, genetic and functional sequence assembly of the barley genome

The International Barley Genome Sequencing Consortium (IBSC)

Gene annotation: M. Spannagl, **M. Pfeifer**, H. Gundlach and K.F.X. Mayer

Transcriptome sequencing and analysis: **M. Pfeifer**, M. Spannagl, P. Hedley, J. Morris, J. Russell, A. Druka, D. Marshall, M. Bayer, D. Swarbreck, D. Sampath, S. Ayling, M. Febrer, M. Caccamo, T. Matsumoto, T. Tanaka, K. Sato, R. P. Wise, T. J. Close, S. Wannamaker, G. J. Muehlbauer, N. Stein, K. F. X. Mayer and R. Waugh

Nature. 491(7426):711–716, 2012.

Integration of whole genome shotgun sequence assemblies with information of genetic and physical map facilitated generating an ordered draft genome sequence and gene annotation for barley (*Hordeum vulgare*). Deep RNA-sequencing provided novel insights in the transcriptome of an agriculturally important member of the Triticeae including, besides expression of protein-coding genes, a high abundance of tissue-dependent alternative splicing, post-transcriptional gene regulation and thousands of novel (non-protein-coding) transcriptional active regions.

Personal contributions: In this project I was mainly responsible for the development and the implementation of a computational workflow for the structural gene annotation of the barley genome sequence assembly by using a multi-tissue RNA-seq data set and public available fl-cDNA sequences. The here implemented gene annotation pipeline was further refined and adapted for the annotation of the bread wheat genome described in **Chapter 4** of my thesis. Furthermore, I conducted the presented transcriptome analysis including quantitative and qualitative characterization of barley gene expression and the investigation of alternative splicing patterns as well as post-transcriptional gene expression regulation. Based on the results observed for the barley genome in this study, my dissertation will further elucidate to which extend the complex transcriptional patterns are also evident in other Triticeae as exemplified by the bread wheat genome (**Chapters 4** and **5**).

2013

MIPS PlantsDB: a database framework for comparative plant genome research

T. Nussbaumer, M. M. Martis, S. K. Roessner, **M. Pfeifer**, K. C. Bader, S. Sharma, H. Gundlach and M. Spannagl

Nucleic Acids Research. 41(D1): D1144–D1151, 2013.

This manuscript describes the web services implemented for data retrieval and data visualization of genomic resources established in the PGSB group for a diverse spectra of plant genomes.

Personal contributions: In this project I contributed to the implementation of the web-based visualization for GenomeZipper results, in particular, for the perennial ryegrass genome.

The perennial ryegrass GenomeZipper – targeted use of genome resources for comparative grass genomics

M. Pfeifer, M. M. Martis, T. Asp, K. F. X. Mayer, T. Lübberstedt, S. Byrne, U. Frei and B. Studer

Plant Physiology. 161(2):571-582, 2013.

In absence of a reference genome sequence, this study applied the GenomeZipper approach to order transcriptome sequence assemblies generated for perennial ryegrass (*Lolium perenne*) by integrating high-density genetic marker maps of *Lolium* and known genome information of the related grasses *Brachypodium*, rice and sorghum. The obtained ordering provided previously unknown insights into the genome architecture of an agricultural and industrial important turfgrass. Moreover, sequence divergence analysis deepened the knowledge of the evolutionary relationship in the Triticeae spanning an evolutionary time frame of approximately 50 million years.

Personal contributions: In this work I was responsible for the comparative genome analysis between perennial ryegrass, barley and available high-quality reference grass genomes. Subsequently, I carried out the GenomeZipper approach, which has been previously established for the analysis of the barley and wheat genomes. This ordering of perennial ryegrass transcriptome sequence assemblies allowed me analysing macro- and micro-synteny relationships between the perennial rye grass genome and the barley genome. Furthermore, I conducted the sequence divergence analysis among related grass genomes.

Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation

J. Jia‡, S. Zhao‡, X. Kong‡, Y. Li‡, G. Zhao‡, W. He‡, R. Appels‡, **M. Pfeifer**, Y. Tao, X. Zhang, R. Jing, C. Zhang, Y. Ma, L. Gao, C. Gao, M. Spannagl, K. F. X. Mayer, D. Li, S. Pan, F. Zheng, Q. Hu, X Xia, J. Li, Q. Liang, J. Chen, T. Wicker, C. Gou, H. Kuang, G. He, Y. Luo, B. Keller, Q. Xia, P. Lu, J. Wang, H. Zou, R. Zhang, J. Xu, J. Gao, C. Middleton, Z. Quan, G. Liu, J. Wang, International Wheat Genome Sequencing Consortium, H. Yang, X. Liu, Z. He, L. Mao and J. Wang

Nature. 496(7443):91-95, 2013.

This paper presents the draft genome sequence for *Aegilops tauschii*, the diploid progenitor genome of the bread wheat D genome. By using high-depth next generation sequencing technology, the authors generated genomic resources providing useful information, e.g., for comparative analysis with the hexaploid wheat genome and related extant diploid genomes that will allow gaining insights into the evolution of the tribe *Triticum*.

Personal contributions: By incorporating genetic marker map information and synteny between *Aegilops tauschii*, barley and related reference grasse genomes, I participated in the anchoring and linear ordering of the predicted genes. In addition, I contributed to the comparative gene family analysis including the computation of orthologous gene relationships and identification of gene families with expanded sizes in the D-genome lineage.

Molecular and immunological characterization of ragweed (*Ambrosia artemisiifolia* L.) pollen after exposure of the plants to elevated ozone over a whole growing season

U. Kanter, W. Heller, J. Durner, J. B. Winkler, M. Engel, H. Behrendt, A. Holzinger, P. Braun, M. Hauser, F. Ferreira, K. F. X. Mayer, **M. Pfeifer** and D. Ernst

PLoS ONE. 8(4):e61518, 2013.

Barley whole exome capture: a tool for genomic research in the genus Hordeum and beyond

M. Mascher, T. A. Richmond, D. J. Gerhardt, A. Himmelbach, L. Clissold, D. Sampath, S. Ayling, B. Steuernagel, **M. Pfeifer**, M. D'Ascenzo, E. D. Akhunov, P. E. Hedley, A. M. Gonzales, P. L. Morrell, B. Kilian, F. R. Blattner, U. Scholz, K. F. X. Mayer, A. J. Flavell, G. J. Muehlbauer, R. Waugh, J. A. Jeddeloh and N. Stein

The Plant Journal. 76(3):494-505, 2013.

The manuscript describes the targeted sequencing of mRNA-coding exons for the barley genome. This approach reduces genomic complexity towards the protein-coding part of the genome and provides a valuable tool, also for the analysis of other Triticeae genomes.

Personal contributions: In this study I contributed to definition of exon sequences for the design of the capturing array on basis of structural transcript assemblies of mapped RNA-seq short reads.

Analysing complex Triticeae genomes – concepts and strategies

M. Spannagl, M. M. Martis, **M. Pfeifer**, T. Nussbaumer and K. F. X. Mayer

Plant Methods. 6;9(1):35, 2013.

This review summarizes the different bioinformatic approaches for the sequence analysis of complex Triticeae genomes.

Personal contributions: I was mainly responsible for the section discussing the orthologous group assembly, which is described in **Chapter 3** of this thesis.

2014

Ragweed (*Ambrosia artemisiifolia*) pollen allergenicity: SuperSAGE transcriptomic analysis upon elevated CO2 and drought stress

A. E. Kelish, F. Zhao, W. Heller, J. Durner, J. B. Winkler, H. Behrendt, C. Traidl-Hoffmann, R. Horres, **M. Pfeifer**, U. Frank and D. Ernst

BMC Plant Biology. 14:176, 2014.

A chromosome-based draft sequence of the hexaploid wheat genome

The International Wheat Genome Sequencing Consortium (IWGSC)

Gene annotation: **M. Pfeifer**, Manuel Spannagl and K. F. X. Mayer

Transcriptome sequencing and expression analysis: **M. Pfeifer** L. Pingault, K. F. X. Mayer and E. Paux

miRNAs: P. Faccioli, M. Colaiacovo, **M. Pfeifer**, A. M. Stanca, H. Budak and L. Cattivelli

Comparative analysis of diploid, tetraploid and hexaploid wheat: **M. Pfeifer**, S. R. Sandve, T. Nussbaumer, K. C. Bader, F. Choulet, C. Feuillet and K. F. X. Mayer

Science. 345(6194):1251788, 2014.

In frame of the International Wheat Genome Sequencing Consortium (IWGSC), chromosome flow cytometry facilitated isolating DNA of individual chromosomes and chromosomes arms of bread wheat. Each chromosome arm was separately sequenced by using high-depth next generation sequencing technology and subsequently assembled *de novo*. The generated chromosome arm survey sequence assembly facilitated a comprehensive analysis of the bread wheat genome revealing high organizational and structural conservation across homoeologous genomes, chromosomes and chromosome arms. Comparative analysis with six extant diploid and tetraploid wheat genomes allowed investigating the phylogenetic relationships and gene family evolution across different *Triticum* genome lineages. These previously unknown genomic resources also enabled elucidating the gene expression patterns with a homoeologous resolution, which revealed a high degree of transcriptional autonomy and no global genome dominance.

Personal contributions: This manuscript constitutes the major publication for the results presented in **Chapter 4** of my thesis and defines the underlying genomic resources utilized in **Chapter 5**. On basis of the generated chromosomal survey sequence assembly, I was mainly responsible for the design and the implementation of an extrinsic gene annotation pipeline and the identification and subsequent classification of high- and low-confidence genes **(Chapter 4)**. On basis of the defined gene sequences and structures I performed a gene expression analysis including five distinct wheat tissues. In particular, I focussed on investigating the transcriptional similarities and differences among homoeologous genes by using similar methods as presented in **Chapter 5** of this thesis. Furthermore, I conducted a comparative sequence analysis of related extant diploid, tetratraplid and hexaploid wheat genomes, implemented the computational workflow for the automated identification of single nucleotide variants and performed the subsequent phylogenetic analysis. Additionally, I was involved in the miRNA analysis and performed the *in silico* prediction of potential gene targets for mature miRNA sequences and the identification of miRNA loci associated with transposable elements.

Ancient hybridizations among the ancestral genomes of bread wheat

T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, **M. Pfeifer**, International Wheat Genome Sequencing Consortium, K. S. Jakobsen, B. H. Wulff, B. Steuernagel, K. F. X. Mayer and O.-A. Olsen

Science. 345(6194):1250092, 2014.

This paper reports a genome-wide comparative analysis investigating the evolutionary relationships among the wheat A-, B- and D-genome lineages. Based on gene annotations for related diploid, tetraploid and hexaploid wheat genomes, the authors estimated the divergence of the A and B genomes approximately seven million years ago. Incongruence in the observed phylogenetic patterns suggested that multiple homoploid and polyploid speciation events shaped the bread wheat genome and indicated that the D-genome lineage resulted from hybridization between the wheat A and B genomes about one to two million years ago.

Personal contributions: In this work I contributed to the underlying genomic resources. Therefore, I applied the reference-guided approach, which is described in **Chapter 4** of my thesis, for the gene annotation of sequence assemblies for the genomes of *Triticum urartu*, *Triticum monococcum*, *Aegilops speltoides*, *Aegilops sharonensis* and *Triticum turgidum*.

Genome interplay in the grain transcriptome of hexaploid bread wheat

M. Pfeifer[‡], K. G. Kugler[‡], S. R. Sandve, B. Zhan, H. Rudi, T. R. Hvidsten, International Wheat Genome Sequencing Consortium, K. F. X. Mayer and O.-A. Olsen

Science. 345(6194):1250091, 2014.

By using RNA-sequencing technology, this study presents a detailed analysis of the bread wheat grain transcriptome for dissected cell types of three different stages in endosperm development. Distinct co-expression clusters were identified, which characterize gene expression in aleurone cells, starchy endosperm and transfer cells during endosperm differentiation and maturation, in which the industrial important characteristics of wheat grains are set. Furthermore, this manuscripts provides previously unknown insights into the contribution of the homoeologous genomes to the grain transcriptome. Genome dominance and genome asymmetry were often found in spatiotemporal co-expression modules and were associated with distinct cellular functions and chromosomal domains. The observations indicated a complex interplay of genetic and epigenetic mechanisms orchestrating gene expression in hexaploid bread wheat grain.

Personal contributions: This study is the major publication for **Chapter 5** of this thesis. On basis of the previously unknown genomic resources, which are described in **Chapter 4**, I updated the existing bread wheat gene annotation by incorporating the novel transcriptome data set generated for wheat endosperm. Therefore, I developed the computational workflow for the mapping of paired-end RNA-seq short reads against the bread wheat draft genome sequence assembly. Furthermore, I conducted an *in silico* evaluation experiment to exclude a bias in the gene expression measurements. I contributed major parts to the statistical analysis of qualitative and quantitative gene expression in developing wheat grain as well as to the interpretation of the observed transcriptional patterns. Additionally, I was mainly responsible for the k-means co-expression cluster analysis of the entire wheat transcriptome as well as the network-based analysis of gene expression regulation for single-copy homoeologous genes. Also, I contributed to the identification and analysis of gene families affecting bread wheat baking quality.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *"To avoid encroaching into already-stressed ecosystems, societies will have to almost* ***double*** *the existing rate of* ***agricultural productivity*** *growth while minimizing the associated environmental damage. This* ***requires dedicated efforts*** *to deploy known but neglected practices,* ***identify crop varieties*** *able to withstand climate shocks, diversify rural livelihoods, improve management of forests, and invest in information systems."*
>
> The World Bank. World Development Report 2010. *World Bank*, page 133, 2009

1.1 Filling the yield gap – new challenges for agricultural research

Worldwide agriculture has been revolutionized in the past 50 years by huge financial investments and international efforts in modern scientific research of crop and livestock production, global distribution of novel technologies, improved infrastructure and systematic market development *(1)*. Breeding of new plant varieties, application of chemicals and fertilizers, irrigation and mechanization of agriculture, significantly increased productivity *(2,3)* and lead to an enormous reduction in global starvation, from approximately one third of the world's population in 1950 *(1)* to one out of eight people in 2012 *(4)*. Due to the indisputable achievements in reducing hunger, USAID Administrator W. S. Gaud designated this era of ground-breaking changes in farming, food processing and management as the "Green Revolution" *(5)*. However, as worldwide population has rapidly doubled within fifty years from three billion in 1960 to seven billion nowadays, the relative decrease in undernourishment is not reflected in terms of absolute numbers *(6,7)*. Most recent estimations of the FAO revealed that still approximately 840 million (mio) people suffered from chronic hunger in 2013 *(4)*. Furthermore, pronounced imbalances exist between different regions and geopolitical areas of the world. Hunger is mainly prevalent in developing countries, where 827 mio people are undernourished (14% of population). On the contrary, less than 5% of population (16 mio people) insufficiently meet daily dietary needs in more developed countries *(4,7)* (Fig. 1.1a). Modernization of agriculture has been successfully implemented in Latin America and the Caribbean as well as Asia and Oceania, which are expected to reach the World Food Summit target that aims at half the absolute number of hungry people of 1990 till 2015 as well as the more challenging Millennium Development Goal that aims at half the proportion of hungry people in the same time frame *(4)*. On the contrary, in Africa the relative decrease in starvation is slowing down and the total number of people suffering from chronic hunger is growing, particularly in Sub-Saharan countries. This trend is also reflected by the significant differences in cereal yield growth during the Green Revolution, which stagnated in Africa, but almost triplicated for other parts of the world *(7)* (Fig. 1.1b).

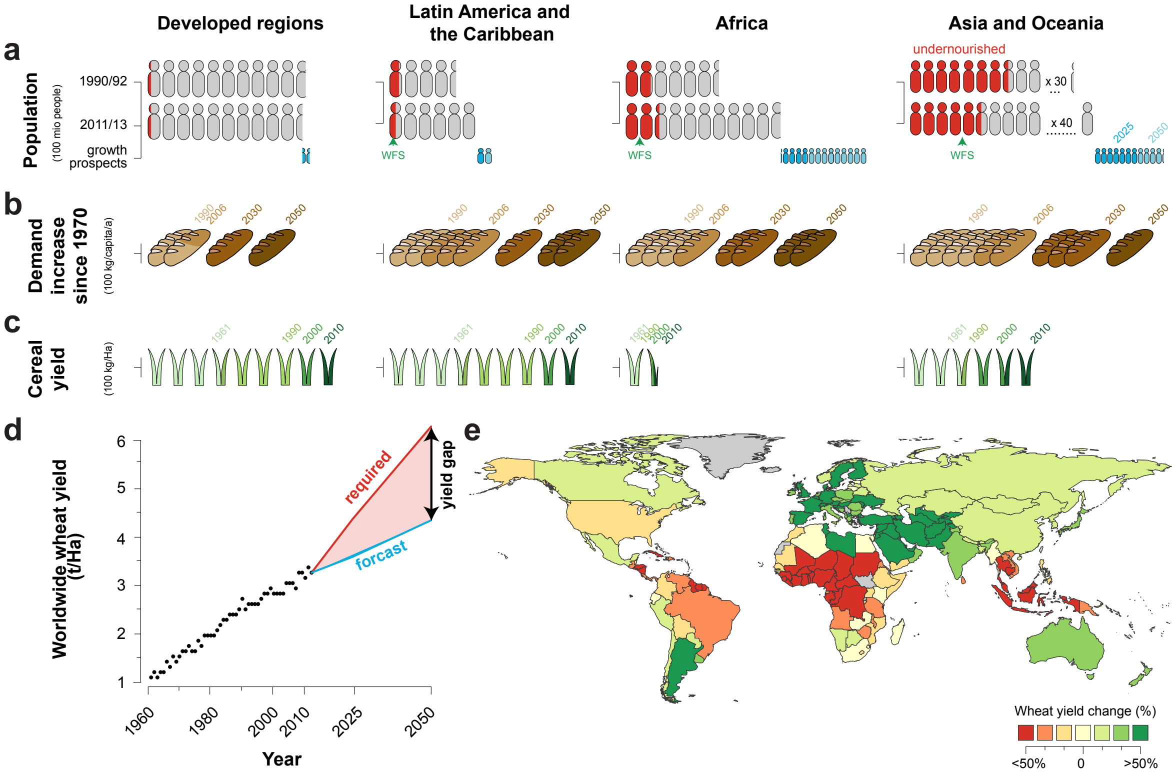


Fig. 1.1. Past successes and new challenges for agricultural research.
Geographic overview of the development of **a,** population size, growth and undernourishment [the regional World Food Summit (WFS) targets aiming at halving the number of undernourished people between 1990/92 and 2015 are indicated by green arrows], **b,** increase in per-capita calorie consumption and **c,** cereal yield growth. **d,** Population growth and changed diets demand for global improvement of agricultural productivity as a "yield gap" emerge by divergence between actual and required increase in cereal yields (red area in inset). **e,** Global statistical predictions indicate strong negative effects of climate change on agroecological conditions especially in Sub-Saharan Africa and South-East Asia in the last quarter of the 21st century. [*Used data sources: Population size: (4,7); Yield growth: (7); Food demand: (8); Estimations of wheat yield predictions: (3); Yield change under climate change: (9,10) (median wheat yields 2070-99 compared to 2005-10; model settings: EPIC, HADGEM2-ES, RCP8.5, SSP2, CO2)*]

Although the Green Revolution constitutes a major improvement, the high number of people suffering from chronic hunger requires further concerted efforts to accomplish global access to protein and energy sufficient for a daily diet *(4)*. Achieving this ambitious goal is impeded by inadequate agroecological capabilities of certain geographic regions, by economic, structural and political constraints of individual countries as well as by limited technology transfer or insufficient international investments in research and development *(4,11,12)*. Furthermore, new challenges will additionally impact on improving the worldwide nutritional situation and will impede ensuring long-term food security. Population growth is predicted to add 3.9 billion people within this century expanding world population to 8.1 billion in 2050 and to 10.9 billion in 2100 *(6)*. While the population in developed countries will stagnate, developing countries rise significantly (3.8 billion) and, in particular, the population size of the least developed countries is assumed to double within the next century (Fig. 1.1a). Moreover, per capita demands for calories and proteins are growing slightly in developed and substantially in developing regions (Fig. 1.1b). Generally, diets are shifting towards increased consumption of livestock products (i.e. meat and dairy) that are generally more resource-intensive to produce *(8,13)*. In combination, population growth and changes in nutritional behaviour cause rising worldwide food demands, which are expected to double until 2050 *(14)*. However, the development of economic needs contrasts with the past and current rates of annual crop yield increase (Fig. 1.1c). Consequently, only approximately 75% of the required calories are predicted to be satisfied in future leading to concerning differences between current and required cereal yield and to the emergence of a "yield gap" *(3)* (Fig. 1.1d). This imbalance indicates that "*the world faces a looming and growing agricultural crisis*" *(3)* and asks for significant improvements in agricultural production *(2,15)*.

However, closing the yield gap by an adequate increase in agricultural production is challenging. Changing climatic conditions "*will depress agricultural yields in many regions, making it harder to meet the world's growing food needs*" (page 133) *(13)*. Recent statistical models predict strong negative effects of warming and increased nitrogen concentration on agriculture, which may cause regional yield losses up to 50% *(9)* (Fig. 1.1e). Furthermore, competing demands of natural resources by food production, bioenergy and biofuel technology or urbanization will increase water and land scarcity *(13)*. At the same time, the established technological inventions and agricultural improvements during the Green Revolution have been associated with environmental damage and pollution *(2)*. For example, extensive use of fertilizes and other chemicals as well as irrigation contributed to pollution of ground water and coastal areas, reduction of biodiversity and increased emission of green house gases *(2,13)*. In conflict with global goals to maintain biodiversity and healthy ecosystems, "*modern agricultural land-use practices may be trading short-term increases in food production for long-term losses in ecosystem services*" *(16)* causing strong negative impacts on agricultural production as well as animal and humankind's life.

Refinement of current proceedings and establishment of new technologies are essential to further reduce worldwide undernourishment and to ensure global food security. The disadvantages of the first Green Revolution and a growing set of other challenges require a second Green Revolution *(17)* and the "*sustainable intensification*" of agricultural productivity *(18)*. Future agriculture has to persistently increase food production independent from environmental perturbations and demographic structures in order to maintain ecosystem services *(15,17,18)*. Besides changes in human mode of behaviour and improvements in organization and infrastructure, biological sciences will play an important role in achieving significant advances in crop yields *(15,19)*. Comprehensive genetic and phenotypic analysis of crop plants provide valuable information to determine and target genes of agricultural importance as well as to identify varieties with favourable traits. Thereby, genome sequencing and analysis are fundamental to unravel an organism's molecular and genetic architecture *(15,18,20–22)*. A complete or draft reference genome sequence of a target crop enables myriad applications including, for example, analysis of evolutionary and phylogenetic relationships, comparative (structural) analysis or the discovery of single nucleotide polymorphism (SNPs) and copy number variation (CNVs) between populations, cultivars or species *(23,24)*. Extensive transcriptome studies investigating plant responses under different environmental conditions can be combined with metabolic and phenotypic data to associate genes with functions, their epigenetic and genetic control mechanisms and underlying regulatory networks and biological pathways *(25,26)*. On-going improvement of high-throughput sequencing technologies and the implementation of bioinformatic approaches integrating different data resources substantially impact genome-wide discovery of key genes and molecular markers *(20,22)*. Studies on a functional and systems level will accelerate conventional or genetic-based plant breeding and will support the identification of "*improved varieties with improved yield and quality, tolerance to unfavourable environmental conditions and resistance to disease*" *(20)*.

1.2 The grasses and the tribe Triticeae

1.2.1 Agricultural and economic importance of the Triticeae

About 10,000 years ago the start of agriculture marked a turning point in history changing humans lifestyle from nomadic hunterer-gatherer to a sedentary, agrarian lifestyle *(27)*. Beginning with the cultivation of barley, emmer wheat and einkorn wheat in the Near East, farming and cultivation of cereals and other food plants expanded across the globe *(27,28)*. The wild progenitors of the firstly cultivated species and their modern varieties belong to the tribe Triticeae, which groups about 300 species including, besides turf and forage grasses, the major cereals *Triticum aestivum* (bread wheat), *Triticum durum* (pasta wheat), *Secale cereale* (rye) and *Hordeum vulgare* (barley) and modern x*Triticosecale* (Triticale) *(29)*. Triticeae are morphologically characterized by open leaf sheaths, membranous sessile spikelets with simple starch grains and hairy ovaries *(30,31)* and are grown in almost all temperate regions around the world *(7)*.

Triticale, barley, rye and wheat provide raw material for myriad industrial applications and livestock feeding and contribute essentially to human diet as staple food of the major civilizations of Europe, West Asia and North Africa *(32)*. In 2012 the Triticeae brought in a collective harvest of approximately 830 mio tonnes and accounted for more than one third of worldwide agriculture land usage *(7)* (Fig. 1.2). Wheat is the most grown Triticeae contributing about one quarter to worldwide crop production (670 mio tonnes in 2012) and generating a trade volumne of more than $200 billion. Barley (132 mio tonnes) and rye (14 mio tonnes) constitute approximately six per cent to global crop production. Despite more than three fold increase in annual yields during the first Green Revolution (Fig. 1.1d), further productivity improvement of agricultural relevant Triticeae species is required to satisfy globally increasing demands in a challenging environmental context *(3,16)*.

1.2.2 Taxonomy and phylogeny of the grasses

The Triticeae is a subgroup of the Poaceae family, one of the largest and ecological dominant families of flowering plants encompassing agricultural important turfgrasses and crops like, for example, millet, sorghum, maize or rice *(34,37)* (Fig. 1.2). The grasses split from a common ancestor approximately 77 mio years ago (mya) *(34)* and have been taxonomically grouped into six major and several smaller subfamilies *(34,38)*. The anomochlooids represent the most early

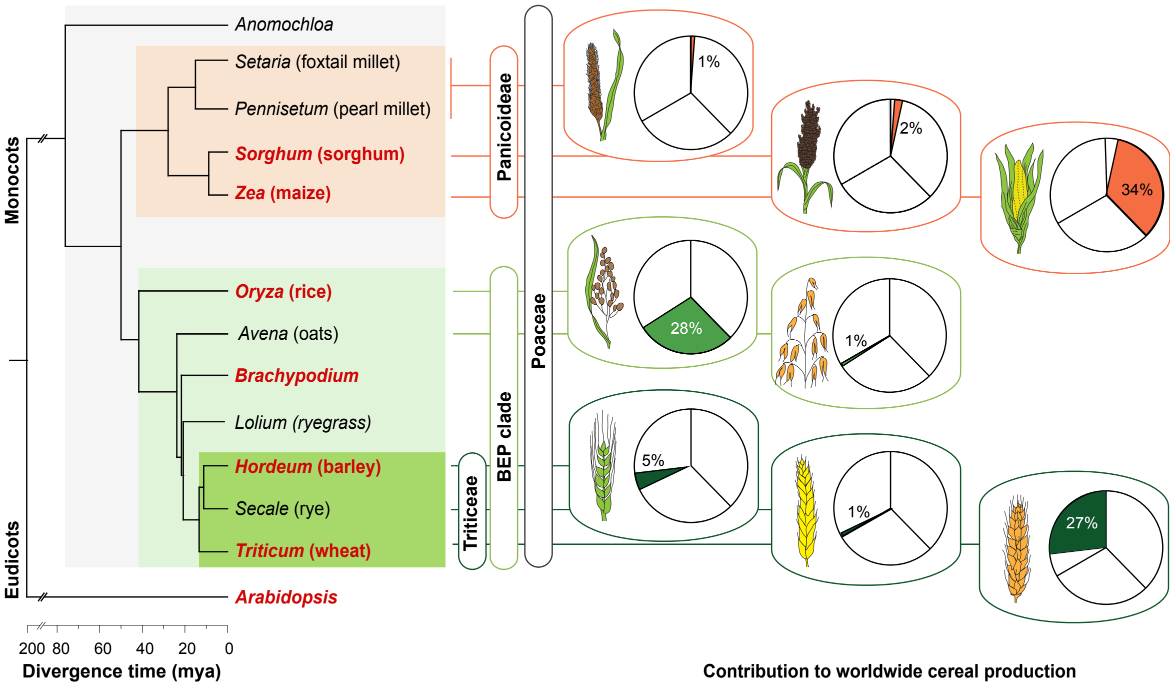


Fig. 1.2. Taxonomy of the Poaceae family and contribution of economically important grasses to worldwide food production.
The dendrogram visualize the general taxonomy of the major economical and scientific important grasses. Estimated divergence times are given in million years ago (mya). Species studied and used for comparative analysis in this thesis are highlighted in red. The relative contribution of each species to worldwide food production in 2012 is shown by the dark portion of the circle diagrams (others: 1%). [*Taxonomy and divergence estimates are based on (33–36). Cereal production statistics are taken from (7).*]

diverged subfamily. The remaining subfamilies are organized in two monophyletic clades, which split approximately 50 mya and group the Panicoideae, Arundinoideae, Chloridoideae and Centothecoideae into the ”PACC-clade” and the Bambusoideae, Ehartoideae and Pooideae into the ”BEP-clade”. Both clades include ecologically important grasses like, for example, millet, maize and sorghum in the PACC-clade and rice, wheat and barley in the BEP clade. About 46 mya the Bambusoideae and Ehartoideae diverged from the Pooideae separating rice (Ehartoideae) from the Pooideae, which include the model grass *Brachypodium*, the oats, the turfgrasses and the Triticeae *(34)*. The oats split first from the Triticeae (approximately 25 mya) *(34)*, followed by *Brachypodium* (approximately 23 mya) *(35)* and the Lolium and Festuca lineages (approximately 22 mya) *(36)*. The Triticeae itself partitioned from a last common ancestor into barley (approximately 13 mya), rye (approximately 11 mya) and wheat *(34)*.

1.2.3 Constitution of grass genomes and syntenic relationships

Genome sizes and chromosome numbers of the grasses are substantially different between and within individual subfamilies, tribes and genera *(29,39)*. This highlights the evolutionary instability and plasticity of plant genomes, which are shaped by frequent changes in the deoxyribonucleic acid (DNA) sequence and chromosomal constitution *(34)*. For example, within the PACC clade the two Panicoideae species *Sorghum bicolor (40)* and *Zea mays (41)* have 10 chromosomes and genome sizes of approximately 0.7 Gb and 2.3 Gb, respectively. Pronounced differences are also present within the BEP clade with 5 chromosomes and approximately 0.4 Gb genome size for *Brachypodium distachyon (42)*, 7 chromosomes and about 5 Gb genome size for (diploid) Triticeae species *(43,44)* or 12 chromosomes and approximately 0.4 Gb genome size for *Oryza sativa (45)*. Often, these differences originated from intra- and interspecies hybridization events and whole genome duplications, which give rise to polyploid organisms with multiplied chromosome numbers. As further discussed in the following (Section 1.3), genome merging and duplication constitute a ”genome shock” *(46)*, which triggers rapid genomic changes in the DNA sequence of particular chromosomes *(47–50)* and consequently may alter chromosome number *(34)*. Moreover, grass genomes vary considerable in the proportion of repetitive DNA sequence, which mainly account for the differences in genome sizes *(51)*. Individual lineages and species have specific rates for amplification and removal of repetitive sequences and distinct signatures to the activity of transposable elements (TEs) *(52)*, a special class of non-genic DNA elements that can replicate, amplify and move to new sites in the genome by a cut-and-paste mechanism or via a RNA intermediate *(53)*.

Despite these large differences, numerous comparative studies that incorporate a diverse spectra of different species revealed high synteny and colinearity in corresponding chromosomal segments of same ancestral origin among the grasses *(54)*. In 1995 Moore *et al. (55)* split the chromosomes of six major grasses (rice, wheat, maize, foxtail millet, sugar cane and sorghum) into 19 linkage blocks, which show significant conservation of gene order, and aligned these segments into concentric ”crop circles” allowing to compare inter-species relationships. The first version of this evolutionary model was based on low-resolution restriction fragment length polymorphism marker maps and was limited in detecting small- and medium-sized structural rearrangements. Accompanied by the improvement of marker maps and DNA sequencing technologies, the crop circle model was further refined by considering additional species and by increasing the resolution using high-density genetic markers, established physical maps and reference genome sequences as well as comprehensive EST collections *(34,35,56–58)*. This allowed inferring the putative arrangement of syntenic blocks in the ancestral grass genome and elucidating the underlying constraints driving speciation and genome evolution. For example, Bolot *et al. (35)* developed an evolutionary scenario, in which the grasses share a whole genome duplication followed by two interchromosomal duplications and fusion events. These events led to an intermediate grass genome consisting of $n = 5 + 5 + 2 = 12$ chromosomes before the split of the PACC and BEP clades. This basic chromosome number is maintained in rice, however, maize and sorghum experienced two additional fusion events ($n = 12 - 2 = 10$), while five subsequent fusion events resulted in an ancestral Triticeae genome of seven chromosomes ($n = 12 - 5 = 7$). Nowadays, the crop-circle model represents a powerful concept facilitating to project positional information from a known grass genome onto a related target genome. This comparative-based approaches have been successfully applied to support molecular genomics and positional cloning *(59)* and improving marker map developments *(60)* or structural genomics *(61–63)* (Section 1.5).

1.2.4 Evolution and phylogeny of cultivated wheats

The genus *Triticum* L. is the economically most important subgroup of the Triticeae accounting for 80% of the tribe's total agricultural productivity *(7)* (Fig. 1.2). It includes wild and cultivated varieties of six species and has been substantially shaped by alloploidization via natural hybridization [Refs. *(64–67)* and references therein] (Fig. 1.3), an evolutionary phenomenon further discussed in the following section of this dissertation. Based on different ploidy levels, the genus *Triticum* has been organized into three sections. The section Monococcon includes two species with diploid genome constitutions (2n=2x=14), *Triticum monococcum* L. (A^mA^m genome) and *Triticum urartu* L. (A^uA^u genome). These diverged about one million years ago and, whereas wild and cultivated forms of *T. monococcum* are known, only cultivated varieties of *T. urartu* exist. Species with tetraploid genomes (2n=4x=21), *Triticum trugidum* L. (AABB genome) and *Triticum timopheevii* Zhuk. (AAGG genome), are grouped into the section Dicoccoideae. For both tetraploids wild and cultivated forms are known. The remaining species, *Triticum aestivum* L. (AABBDD genome) and *Triticum zukovsky* L. (AAAAGG genome), group in the section Triticum, have hexaploid genome constitutions (2n=6x=42). For both hexaploid wheat genomes only cultivated forms have been reported so far.

The complex structure of the genus originated in multiple, independent hybridization events *(64,65,67)*. About 0.8 mya, incidental hybridization between wild *T. urartu* and diploid species belonging to the *Aegilops* genus, believed to be related to *Aegilops speltoides* (2n=2x=14; SS genome), resulted in tetraploid *T. turgidum* (AABB genome) and *T. timopheevii* (AAGG genome).

Subsequently, with the beginning of farming, these tetraploid varieties were cultivated and free-threshing forms evolved from the hulled genotypes, having soft glumes and being easier to harvest *(65,68)*. Simultaneously, hexaploid *T. aestivum* [bread wheat (AABBDD genome)] and *T. zukovsky* (AAAAGG genome) emerged from hybridizations between the two domesticated tetraploid wheat genomes with the wild diploid species *Aegilops tauschii* (2n=2x=14; DD genome) and with the cultivated form of *T. monococcum*, respectively.

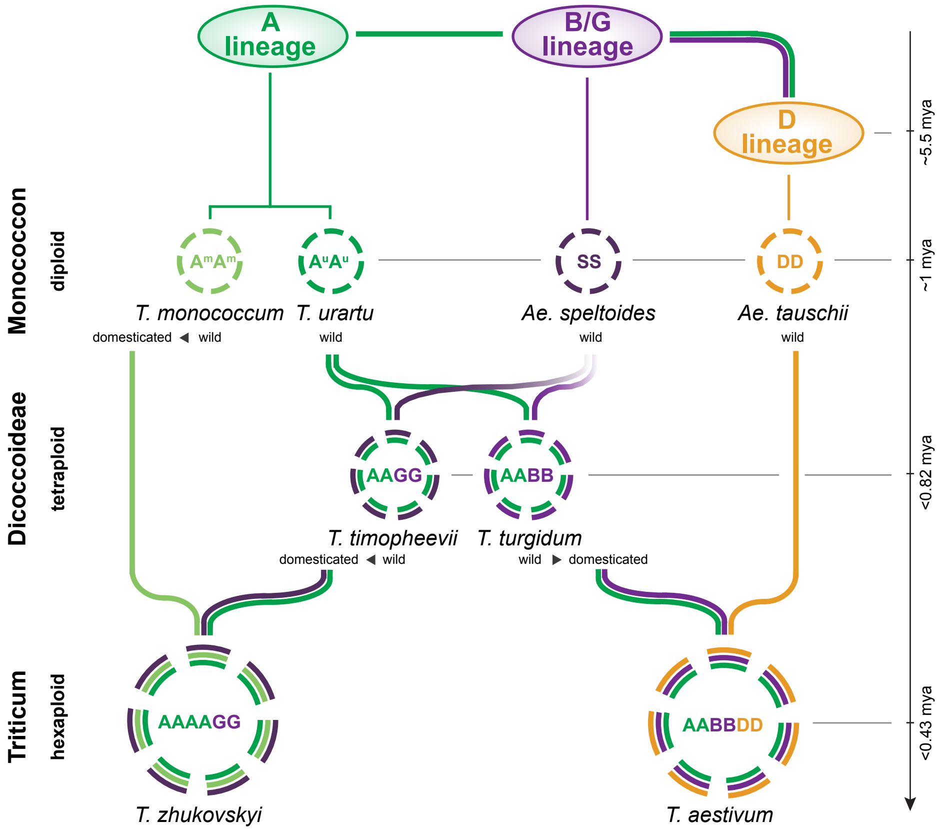


Fig. 1.3. Evolution of di-, tetra- and hexapolyploid wheat genomes of the genus *Triticum*.
For each species nomenclature within circles provide a schematic representation of the genome constitution. Thin lines indicate linear evolution of ancestral diploid genomes in the A-, B-/G- and D-genome lineages, while bold lines visualize hybridization events resulting in tetra and hexapolyploids. [*Phylogeny and time estimates are based on (64,65,69,70).*]

Moreover, recent comparative analysis utilizing molecular data and genomic sequence resources suggested incongruent, reticulate evolution of the different Triticeae lineages and reported introgressive events like, for example, hybridization, gene flow or horizontal gene transfer *(71–73)*. Based on genome resources established within this thesis, Marcussen *et al. (69)* conducted a genome-wide analysis of the evolutionary relationships between the A, B and D genomes of bread wheat and related diploid genomes. Inconsistent patterns across gene trees were observed with a higher frequency of *B(A,D)* and *A(B,D)* tree topologies suggesting ancient, inter-lineage hybridization between species of the A and B genome lineages and homoploid hybrid speciation of the diploid wheat D genome progenitor approximately 5.5 mya.

1.3 Formation and implications of polyploidy

1.3.1 Formation and incidence of polyploidy

Polyploid organisms have genomes with an increased number of basic chromosomes *(74,75)*, an evolutionary phenomenon common to many eukaryotes including plants *(76,77)*, fish *(78)*, vertebrates *(79)* and fungi *(80)*. Cells with multiplied genome sets derive from somatic doubling or the incidental formation and fusion of gametes that contain more than one set of chromosomes *(74)*. Based on the type and origin of the multiplied chromosome sets, Kihara and Ono *(81)* proposed a classification of polyploids into ”autopolyploids” and ”allopolyploids” *(74)* (Fig. 1.4a). The former type, autopolyploids, arise from doubling the chromosomes of a diploid genome by, for example, fusion of two diploid gametes. On the contrary, the latter type, allopolyploids, result from the merger of chromosome sets of different genomes, for example, by interspecific hybridization of two haploid gametes followed by chromosome doubling, of two diploid gametes or of gametes from distinct autopolyploids. However, with higher similarity of the parents, the distinction between auto- and allopolyploidy becomes blurred *(82)*.

Most eucaryotic genomes are innate polyploids and experienced one or more whole genome duplication (WGD) events *(76–80)*. Generally, plants have a relatively high polyploid tolerance and formation rate (approximately one formation per 100,000 individuals) *(74)*. Evidences of polyploid ancestry have been found for more than 70% of flowering plants *(84)* and many species, including *Arabidopsis (85)*, maize *(86)* or rice *(87)*, have secondarily diploidized

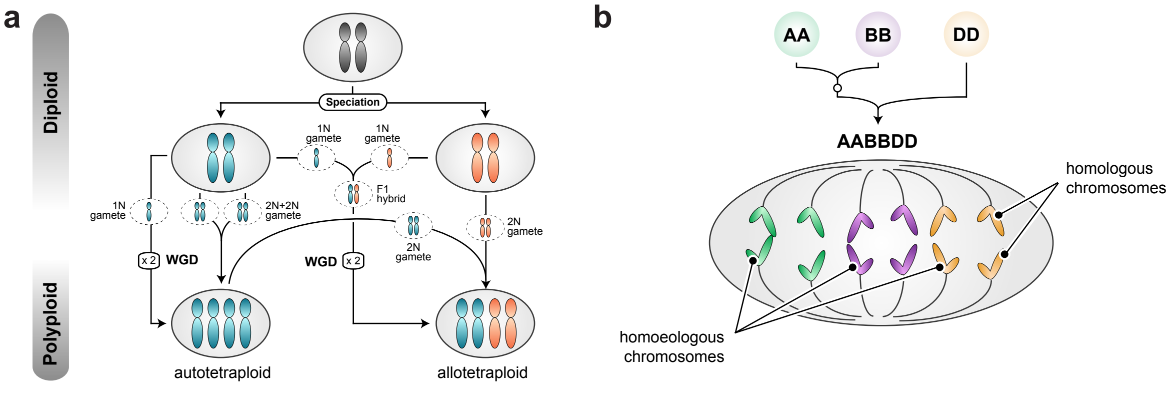


Fig. 1.4. Evolutionary scenarios of the formation of polyploids and bivalent pairing of chromosomes during meiosis.
a, Possible evolutionary alterations resulting in the transition of diploid species to allo- and autotetrapolyploid organisms. Hybridization events are visualized by fusing lines and whole genome duplication (WGD) events are marked by ”2x”. Dashed lines depict the haploid forms of a diploid or tetraploid organism. For simplicity not all possible path are shown. **b,** Schematic illustration of bivalent chromosomal pairing during meiosis exemplified for allohexaploid *T. aestivum* (bread wheat). Homoeologous chromosomes derived from different parental genomes are distinguished and, in a diploid-like behaviour, only identical (homologous) chromosomes pair. For simplicity only two of the seven homoeologous chromosomes are shown for each genome. [*Manually adapted on basis of schematic illustrations in (82,83).*]

genomes and returned to a diploid genome constitution after the polyploidization event. Several plant lineages encompass di- and polyploid members, formed by inter-specific hybridization between genomes from the same genera [e.g. *Brassica (47)*, *Gossypium (88)* and *Triticum (67)*] or between genomes from different taxa [e.g. *Triticum* and *Hordeum (89)*]. Complex evolutionary patterns and multiple subsequent hybridization events have contributed to speciation like, for example, in the genus *Triticum* with allohexaploid bread wheat (Fig. 1.3).

Auto- and allopolyploids usually differ in chromosomal pairing during meiosis *(82)*. Multiplied chromosomes of each type are present in autopolyploids as identical, so called "homologous", copies, which usually exhibit multivalent pairing. Contrarily, in allopolyploids the corresponding "homoeologous" chromosomes, which are similar but differ in their parental origin, are distinguished and pair as bivalents. In principle, this mimics a diploid-like behaviour during meiosis preventing inter-genomic recombination. For example, individual chromosomes of the A, B and D genomes of hexaploid bread wheat pair only with their corresponding homolog (i.e. 1A and 1A) and never with the homoeologous counterparts (i.e. nor 1A and 1B, nor 1A and 1D, nor 1B and 1D) (Fig. 1.4b). In wheat, such accurate and efficient bivalent pairing is controlled by two independent systems *(90)*. The *Ph1* (*Pairing homoeologous 1*) gene constitutes a genetic control instance, which facilitates distinguishing between chromosomes of different origin while allows for intragenomic pairing *(91,92)*. *Ph1* has been suggested to be involved in controlling the interactions between centromers and microtubles and to affect sister chromatid cohesion through alterations in the heterochromatin decondensation. Complementary, a second control instance distinguish homoeologous chromosomes due to physical differences set by rapid alterations in DNA sequence *(48,49)*. These cause genome down-sizing and re-patterning and generate unique chromosomal signatures enabling the distinction between homoeologous chromosome during meiosis (Section 1.3.3).

1.3.2 Polyploidy affects plant vigour and phenotype

Some of the major agriculturally important plants are ancient or innate polyploids. No wild form of allohexaploid wheat is known so far and, thus, one of the major crops is assumed to be a polyploid product of human farming and domestication *(93)* (Fig. 1.3). The worldwide distribution and agricultural importance of polyploids mirrors beneficial effects of genome doubling or merger, which often result in heterosis, i.e. more vigorous characteristics of a polyploid species compared to its parents *(82,94)*. For example, superior traits and phenotypes could arise by increased heterozygosity, which normally declines in the generation of diploid F1 hybrids, but is maintained in allopolyploid progeny. Furthermore, duplications of homoeologous chromosomes lead to a redundant gene pool, which has protective effects by masking recessive alleles derived from one parental genome or allows increased diversification by evolving of one homoeologous gene, while another copy still exerts the innate gene functions.

Morphologically, cell volume increases with ploidy level changing cell structure and the arrangement of cellular components *(95)*. Such changes in the geometric relationships within cells may affect the cellular biochemical mechanisms and, for example, trigger changes in enzyme activity *(96)*, metabolism rates *(97)* and cell-surface related processes *(98)*. However, the increase in cell size may not necessarily result in changes of body size *(82,99,100)* and most phenotypic variations have been suggested to be caused by genetic or epigenetic mechanisms *(83)*, which trigger up- or down-regulation of genes involved in energy and starch metabolism, growth and flowering pathways *(101,102)*.

1.3.3 Implications of polyploidy on the bread wheat genome and transcriptome

Genome doubling or merger constitute a ”*genome shock*” *(46)*, which is accompanied by severe changes in the cellular architecture *(95)* and irregulations during cell division *(104)* as well as causes novel intra- and intergenome interactions and altered regulatory mechanisms *(105–108)* (Fig. 1.5). Based on research in the Triticeae and the wheat lineage, Feldman and Levy *(90)* distinguished between ”revolutionary changes”, which are initialized instantaneously during or immediately after polyploid formation, and ”evolutionary changes”, which occur during the polyploid’s evolution. Extensive chromosomal re-patterning and massive changes in the DNA composition of the inherited chromosomes include often loss of coding and noncoding DNA sequences *(48,49)* or activation of transposable elements *(109)*. Those revolutionary changes constitute improved fertility and polyploid establishment, ensuring intragenomic (bivalent) pairing and rapid elimination of detrimental genetic intra-genomic incompatibilities, whilst mid- and long-term evolutionary changes tend to contribute to beneficial environmental adaption and improved fitness *(90)*. However, immediately induced genomic changes and the following evolutionary processes may lead

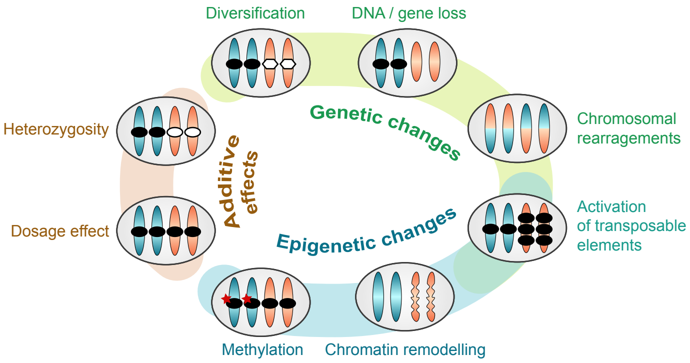


Fig. 1.5. Possible mechanisms affecting the fate of homoeologous genes in polyploid genomes. Polyploid formation triggers alterations in the genomic and transcriptional landscape of the inherited genomes, which has substantial implications on the fate of duplicated genes. Beneficial additive effects may result from an extra gene dosage and increased or heterozygosity. Genetic changes include chromosomal rearrangements, loss of non-coding and coding DNA sequences or other sequence changes. Mutations in the coding sequences or in the regulatory elements may cause functional diversification of homoeologous genes. Epigenetic changes involve chromatin remodelling and alterations of methylation patterns, which provides flexible control mechanism for the transcriptional activity of homoeologous genes. [*Manually adapted on basis of a schematic illustration in (103).*]

to genome fractionation and structural diploidization, which might be biased towards preferential retention and losses of genes from either parental origin *(110,111)*.

Several possible mechanisms could effect the fate of homoeologous genes in a polyploid genome [Refs. *(83,90,103,112)* and references therein]. On the one hand, homoeologous gene copies may be retained in the genome as an additional gene dosage might provide advantageous effects on some gene functions or beneficial intergenomic interactions are established by different regulation of homoeologous genes. On the other hand, gene duplications may disturb the cellular products and pathways with negative implications on the polyploid's fitness. This requires adequate mechanisms compensating for detrimental effects and orchestrating polyploid gene regulation. Genomic changes through the accumulation of mutations, evolution of *cis*-regulatory elements or changes in DNA sequence cause the removal, inactivation or pseudogenization of one gene copy, but may also trigger functional divergence of homoeologs (i.e. sub- and neofunctionalization). In addition, epigenetic mechanisms like, for instance, alterations in the methylation patterns of homoeologous genes, may contribute to the evolutionary advantages of polyploids as flexible, potentially reversible markings allowing development of novel traits and faster response to changed environmental conditions.

Differences in the relative expression levels have been observed for a substantial fraction of duplicated genes in various polyploids including allotetraploid cotton *(113–116)*, *Arabidopsis (101,117,118)* or *Tragopogon miscellus (119)* as well as synthetic and natural wheat allopolyploids *(120–125)*. Immediately after genome merger, Kashkush *et al. (121)* and He *et al. (122)* found approximately 5% of genes with altered gene expression in synthetic allotetra- and allohexaploid wheat genomes. Analysis of genome-specific nucleotide polymorphisms, which discriminated between homoeologous cDNA sequences, revealed 12% (of 90 analysed genes) *(123)* and 27% (of 236 analysed genes) *(124)* of genes to be homoeologous-specific silenced in natural wheat polyploids. Consistent with studies in other polyploids *(113,126,127)*, notably, the higher percentage of silenced genes in established polyploids suggested increasing impact of polyploid evolution on gene expression of homoeologous genes over time. Moreover, homoeologous genes have been found to be regulated differently in different wheat organs. By investigating the gene expression of 79 genes in ten tissues of hexaploid wheat, Mochida *et al. (123)* observed for no gene predominant expression from a specific genome in all tissues. Only 15 genes from each genome (19%) were uniformly expressed across tissues, while the remaining homoeologs exhibited preferentially expression of one genome in at least one tissue. Similar observations have been made by Bottley *et al. (124)* in wheat or by Adams *et al. (126)* in cotton. So far, however, current knowledge is restricted to a limited number of genes and the underlying regulatory mechanisms have not been fully resolved yet. Preferential expression of homoeologous genes in certain tissues might be already established in the parental genomes and the responsible regulatory networks inherited by and maintained in the polyploid hybrid. Alternatively, tissue-specific expression may also suggest functional divergence and indicate sub- and neofunctionalization of homoeologs *(128–130)*, caused by alterations in the genetic regulatory elements (e.g. mutations in transcription binding sites) *(131)* and by epigenetic modifications *(132)*.

1.3.4 Genome asymmetry and homoeolog expression bias

Genome asymmetry and homoeolog expression bias, i.e. favourable expression of homoeologous genes, is common to many polyploids. A genome-wide bias towards one compound of the polyploid genome has been shown for allotetraploid cotton *(113,114,133)*, paleoploid maize *(111)*, mesoploid *Brassica (134)* and synthetic *Arabidopsis* polyploids *(101,118,127)*. On the contrary, no overall transcriptional dominance for one genome has been evident from small-scale studies in allohexaploid bread wheat *(123,124)*. However, genome asymmetry has also been observed in the control of distinct agricultural and industrial important traits. Thereby, the individual genomes contribute differentially to individual morphological, physiological and molecular characteristics: The A genome has been associated with morphological characteristics including plant and spike growth and determining non-brittle rachis *(68)*. As investigated and summarized by Feldman *et al. (135)*, tolerance to environmental challenging conditions and responses to abotic and biotic stresses are more contributed by the B and D genomes, which exclusively contain genes responsible for boron tolerance, iron deficiency or low cadmium uptake (B genome) or aluminium and salinity (D genome). Wheat baking quality and controlling the production of starch and storage proteins during grain filling, which have been associated with B and D genome encoded genes *(136)*.

1.4 Genome and transcriptome sequencing technologies

1.4.1 The ”evolution” of sequencing technologies

The field of genome and transcriptome analysis has dramatically changed during the last two decades. Since the release of the first plant genome sequence for *Arabidopsis thaliana* in 2000 *(137)*, which was generated by using automated Sanger sequencing *(138–140)*, novel high-throughput genome and transcriptome sequencing techniques have evolved to meet the increasing demand in sequence information *(141,142)*. These ”next generation sequencing” (NGS) methods allow cost-efficient generation of comprehensive genome and transcriptome sequences resources for myriad applications in fundamental, industrial or medical research. Ongoing improvements and the advantageous combination of first and second generation sequencing methods have accelerated the release of draft or complete reference genome sequences for many species (Fig. 1.6). These have built the basis for accelerated crop improvement *(143)* by identification of genes and their function allowing to make further associations between genotypes *(24,144)* and phenotypic variations *(23)*.

First generation sequencing: classical DNA sequencing technologies

The history of DNA sequencing methods is distinguished into three epochs. Beginning in the 1970s, DNA sequencing technologies of the first generation were developed utilizing polyacrylamide gel electrophoresis to separate fragments of different sizes generated from a target DNA template *(138,139,145)*. Therefore, each fragment is terminally primed with a radioactive or fluorescence markers that specify each nucleotide type. Spatial ordering by fragment size allows to read the encoded DNA sequence from the emitted signal. Simultaneously, two methods were developed, mainly differing in the methods for cutting the DNA template and for labelling those. Maxam and Gilbert *(145)* applies a series of chemical reactions to cleave a terminally radiolabeled DNA fragment at distinct base positions and infers the sequence along the electrophoretic banding patterns. On the contrary, the method developed by Sanger and Coulson *(138,139)* applies DNA synthesis with polymerase reactions to generate primed fragments from a DNA template. Each nucleotide type is replaced by fluorescently labelled, chain-terminating analogs in one reaction. These impede chain elongation and cause disruption of the DNA polymerase reaction, thus, generating DNA fragments of any size marked by the corresponding termination nucleotide. The DNA fragments are spatially separated by gel-electrophoresis and the DNA sequence inferred from the combination of the four parallel dideoxy reactions.

Due to less laborious sample preparations, reduced chemical requirements and increased sequence read length, the Sanger method has became the favourably used sequencing strategy. Further technological improvements like, for example, capillary gel electrophoresis *(146)* or sequencing of complementary DNA (cDNA) sequences to obtain expressed sequence tags (ESTs) *(147)*, automation of Sanger sequencing *(140,148)* as well as advances in computational data management and bioinformatic analysis *(149)* have contributed to the exponential growth of nucleotide sequence data bases *(150)*. This data increase mirrors also the valuably of genome and transcriptome sequence data for myriad applications.

Second generation sequencing: state of the art technologies

Sanger-based DNA sequences are of high-quality and have approximately 1,000 base pairs (bp) length and less than 0.001% error rate *(151)*. However, relatively elaborative sample preparation and high costs (approximately \$500 per megabase) are confronted with an increasing demand for comprehensive genomic and transcriptome data sets. This has driven the development of alternative sequencing methods, known as second or next generation sequencing technologies *(142,151–153)*. Various technologies that differ in template preparation, sequencing biochemistry and imaging procedures have been implemented including micro-electrophoretic methods, sequencing by hybridization, real-time sequencing and cyclic-array sequencing. In this thesis the utilized genome and transcriptome sequence resources were generated with two major commercial NGS implementations, Roche 454 pyrosequencing *(154)* (Roche Applied Science, Basel, Switzerland) and Solexa/Illumina sequencing technology *(155)* (Illumina Inc., San Diego, California, USA). The two methods differ substantially in the applied chemicals and underlying biochemical processes, however, they share with ”cyclic-array sequencing” the basic technological principle *(142,153)*. Prior to sequencing the DNA sample is randomly fragmented into smaller pieces, which are ligated to DNA primers attached to a support or solid surfaces and amplified. This allows parallelisation of the sequencing reactions for billions of identical DNA fragments in a series of sequencing cycles to read the DNA sequences from the superimposition of the observed imaging signals.

Emulsion polymerase chain reactions / pyrosequencing – 454 pyrosequencing technology, implemented in the Roche 454 Genome Sequencers (Roche Applied Science, Basel, Switzerland), uses emulsion polymerase chain reactions (emPCR) to amplify DNA fragments *(142,154)*. DNA fragments are bound to primer-coated beads forming complexes, which are enclosed in droplets and spatially separated in an oil-aqueous emulsion. Within each compound individual polymerase chain reactions (PCR) reactions are performed to amplify the DNA fragments. Thereafter, the emPCR beads are dissociated and partitioned into millions of wells that are located on a PicoTiterPlate. Supplementary chemistries, including a DNA primer, sulphurylase, luciferase and apyrase, are added to the sequencing reaction. In a series of cycles dideoxynucleotides are disposed across the wells facilitating chain elongation by the DNA polymerase and triggering a chemical reaction resulting in light emission by the luciferase. The intensity of the generated light signal is measured by a high-resolution charge-coupled device camera allowing to read the complementary nucleotide encoded by the DNA template.

Solid-phase amplification / cyclic reversible termination – In Solexa/Illumina sequencing, implemented in the Illumina Genome Analyzers (Illumina Inc., San Diego, California, USA), solid-phase amplification and cyclic reversible termination are applied to read the sequence of a DNA template *(142,155)*. During template preparation the DNA fragments are bound to 5'-primers on a glass slide. Denaturation breaks the double stranded templates and the single stranded DNA strands bind to adjacent 3'-primers to form bridges between primer pairs. These newly formed amplicons are extended by polymerases forming double-stranded bridges. Subsequent denaturation result in two covalently bound single stranded DNA copies. The cyclical repetition of this process generates millions of template copies, which are spatially separated in template clusters on the solid surface of the flow cell. Then, the sequencing reaction is initiated by hybridization of the free ends of the DNA templates with sequencing primers. Fluorescently labelled dideoxynucleotides are added and, complementary to the free position of along the DNA template, a single nucleotide is captured and attached by the DNA polymerase. The chain elongation process terminates and, after removal of the remaining nucleotides, the identity of the bound nucleotide is determined by the emitted fluorescence signal. The chemical constraint blocking DNA polymerase activity are released and the sequencing cycle is repeated.

The two sequencing technologies vary considerably in instrument costs (approximately $500,000 for 454 pyrosequencing and $540,000 for Solexa/Illumina), per megabase sequencing costs (approximately $60 and $2), run time (10 to 23 h and 5 to 65 h), throughput per run (700 Mb and up to 1.8 Tb) and sequencing read length [up to 1,000 bp (mode 700 bp) and maximum 2 x 300 bp] *(142,156,157)*. Each technology has its individual advantages and limitations that significantly influence the downstream analysis. For example, longer reads generated with 454 pyrosequencing technology may improve mapping of repetitive genome regions or facilitate to distinguish between highly similar homoeologous sequence copies obtained for a polyploid genome. Contrarily, Solexa/Illumina sequencing allows cost-efficient high-depth genome and transcriptome profiling to quantify messenger RNA abundances as well as to detect sequence variations between individuals of a population, cultivars or species.

Third generation sequencing: the future of DNA sequencing

This thesis as well as most current genome and transcriptome sequencing projects rely on data resources obtained with NGS technologies. However, third generation sequencing (TGS) platforms became available recently. These aim at an increase in read length, the removal of amplification artefacts, a simplified sample preparation and a decrease in run time *(153,158)*. TGS mainly implement strategies for identification of nucleotides from unmodified DNA strands via physical recognition with nano pores *(159)*, single molecule real time sequencing *(160)* or direct imaging of DNA with electron microscopy *(161)*. Especially improved read length of several thousand kilobases, will be valuable for future genome analysis and will improve the assembly of complex, repeat-rich and polyploid genomes *(162)*.

1.4.2 Bioinformatics – a key discipline for genome and transcriptome analysis

The automation of DNA sequencing and emergence of high-throughput NGS technologies required computation approaches for efficient organization and interpretation of an overwhelming amount of data *(142)*. Thus, bioinformatics developed rapidly into an important, interdisciplinary scientific area with key responsibilities in sequence-driven biological research. The following paragraphs briefly exemplify some fundamental bioinformatic challenges concerned with this thesis. These include the assembly of individual sequencing reads by using *de novo* or reference-guided approaches and the investigation of transcriptome responses based on high-depth cDNA sequencing (RNA-seq).

Basic principles of *de novo* sequence assembly algorithms

As the DNA templates are usually (much) longer than obtained sequencing reads, contiguous sequences have to be reconstructed by using reference-guided methods or *de novo* assembly strategies. Whereas algorithms of the former type require prior genome information and are based on alignments of reads against a known reference genome sequence, *de novo* approaches reconstruct the original sequence on basis of mutually completing sequence information among reads *(163)*. This is a computationally complex and and resource-intensive task, especially, for NGS data sets with shorter and manifold higher sequencing depth compared to Sanger-based resources. Additionally, missing parts in the generated sequence data or sequencing errors complicate the computation of overlaps between reads. However, various approaches and software tools have been developed to assemble reads into contiguous sequences (”contigs”) *(154,164,165)*. Due to underlying algorithmic principles, these could usually be categorized into ”overlap - layout - consensus” (OLC) assemblers and *de Bruijn* graph assemblers *(166,167)*. OLC assemblers construct an overlap graph connecting reads with shared sequences identified by all-vs-all pairwise sequence alignments. This graph is layouted and a consensus sequence inferred by merging connected reads. On the contrary, *de Bruijn* graph assemblers split the reads into k-mers and connect those with (k-1) identical sequences. This translates into a directed graph structure facilitating to compute the minimal ”Hamilton cycle” of the graph, which is the path going exactly once to each node and ending at the starting node. Accordingly to this ordering, k-mers are concatenated to reconstruct the original sequence.

Different assembly methods are differently suited for individual sequencing technologies and read types *(167)*. OLC algorithms favour longer reads to reliable detect overlaps and, thus, are the methods-of-choice for the assembly of reads generated with Sanger sequencing or 454 pyrosequencing. As the computational complexity in determining pairwise alignments among reads increases with genome size and with sequencing depth, *de Bruijn* graph assemblers become more attractive for Solexa/Illumina sequences that usually are produced in high depth to compensate for shorter read length. Thereby, large-scale evaluation studies have shown enormous variations in performance and correctness between current assembly software packages *(168–170)*. Besides a strong impact of data quality and library design, in particular, the specific characteristics and sequence composition of the analysed genome (Section 1.5) itself significantly influence assembly quality. Therefore, the reconstruction of most plant genomes from sequences is a general challenge for *de novo* strategies due to large genome sizes and high amounts of repetitive sequences *(171,172)*.

Next generation short read alignment

With increasing availability of (draft) reference genome sequences (Fig. 1.6), cost, time and effort considerations make the analysis of NGS data via the alignment of obtained reads against a sequence of the target genome or a closely related species interesting for various biological applications *(142,173)*. This approach provides nucleotide-level resolution information usable, for example, in reference-guided assemblies of closely related genomes, in re-sequencing projects investigating genetic variations in populations or between species, in structural annotation of genes and transcripts, or in expression studies. Similar to *de novo* assembly approaches, technical factors (e.g. billions of short reads, sequencing errors and gapped or spliced alignments) and biological aspects (e.g. large genome sizes, genetic variation, repetitive non-coding sequences, duplicated sequences) make alignment approaches computationally difficult. To overcome these challenges, special algorithms and software packages, so called ”short-read aligners”, have been designed *(174–177)*. Different in implementation and application, however, all these programs apply indexing-strategies, which increase time- and memory efficiency and allow fast identification of shorter sequences in a large DNA sequence.

For example, this thesis utilizes Bowtie *(174)*, one of the most frequently applied short-read aligners for the mapping of RNA-seq reads. Prior to the search phase, Bowtie creates a memory-efficient representation of the reference genome sequence by computing an index with the Burrows-Wheeler Transformation (BWT) *(178)*. Then, query sequences are mapped in a character-by-character search aiming at narrow the set of potential alignment positions in the BWT index. To account for mismatches caused by sequencing errors or genetic variations between reference and query, Bowtie implements a backtracking algorithm identifying and substituting the minimum number of positions that do no exact match in the BWT index. Despite on-going improvement of NGS data quality and associated high-performance software packages, balancing accurate alignment of reads and computational efficiency remains challenging, demands for further technological improvements and requires a cautious interpretation of the obtained results *(179)*.

Assembly of gene and transcript structures and quantitative expression analysis

NGS technologies not only revolutionized genomic studies. Furthermore, monitoring messenger RNA (mRNA) abundances by high-depth cDNA sequencing (RNA-seq) facilitates sensitive and accurate analysis of transcriptional landscapes *(180–182)*. In addition to quantitative and qualitative expression analysis, RNA-seq is also particularly valuable for the identification of genes and the structural annotation of transcripts. The nucleotide-level resolution enables to detect and investigate alternative expression for distinct splicing variants, which has been recently shown to contribute important cellular functions in both, mammals *(183)* and plants *(184)*.

Besides general difficulties in handling NGS short reads, computational analysis of RNA-seq reads is faced with substantial variances in sequencing depth caused by differences in expression levels and difficulties in the unambiguous assignment of reads to individual exons and distinct splicing variants *(185)*. In absence of a suitable reference genome sequence, RNA-seq reads can be assembled *de novo* into partial or full-length transcript sequences following similar approaches as described above. However, reference-guided approaches utilizing alignment of RNA-seq reads against a known reference genome sequence are favourably applied for reconstructing gene and transcript structures as well as quantifying expression levels. These approaches demand for less sequencing depth and computational requirements, are more sensitive and accurate especially for annotation of low abundant transcriptions as well as allow the detailed structural definition with exon/intron boundaries of transcripts *(185)*.

With TopHat *(186)* and Cufflinks *(181)*, Trapnell *et al.* implemented two open-source software packages that nowadays belong to the major computational workflows for the alignment of short RNA-seq reads against a reference genome sequences (TopHat) and subsequent transcript reconstruction and expression quantification (Cufflinks). As short RNA-seq reads are generated on basis of processed mRNA (i.e. introns have already been removed by the spliceosome), the alignment requires specialized algorithms that consider also reads spanning exon-exon junctions *(173)*. While some tools apply machine learning approaches to identify reads bridging introns, but rely and are trained on known structural gene annotations *(187)*, TopHat applies an incremental alignment strategy. First, reads are detected that fall entirely into single exons and, secondly, potential splice sites between introns are determined and multi exon-spanning reads mapped. Based on these alignments Cufflinks implements a graph-based representation that connects mutually compatible reads, which overlap at same exon-exon boundaries and belong to same transcripts. This strategy enables distinguishing between incompatible reads, which align to different exons and consequently origin from different transcripts. Accordingly to Dilworth's Theorem *(188)* the minimum number of paths through this overlap graph including each node at least once explains all incompatibilities among fragments, an assumption, which allows Cufflinks reconstructing a minimum set of transcript structures for each gene loci in polynomial runtime *(181)*. Moreover, Cufflinks implements a statistical model to estimate the transcript abundances as a function that best explains the observed transcript coverages by compatible RNA-seq fragments.

1.5 Plant genome sequencing and analysis

1.5.1 Progress in plant genome sequencing

The field of plant biology and genome research was revolutionized with the completion of the *A. thaliana* genome sequence in 2000 *(137)*. The authors applied classical DNA sequencing utilizing hybridization and PCR-based approaches to sequence individual bacterial artificial chromosome (BAC) clones. Considering the arrangement along the minimal tiling path, overlapping BAC sequences were merged and contiguous DNA sequences reconstructed for each chromosome arm. Such BAC-based physical mapping strategies using Sanger sequencing technology are time- and cost-expensive, however, a high-quality genome sequence of the first genome of an agroecological cereal, *Oryza sativa* (rice), was released in 2005 *(45)*. Thereafter, alternative sequencing methods were developed and ”whole genome shotgun” (WGS) strategies applied to generate large collections of random DNA fragments resampling the entire genome. These libraries are produced with a high genome coverage. The redundancy in the obtained genomic data set allows assembling overlapping sequencing reads into contigs. However, the subsequent ordering of the obtained contigs remains challenging and requires, for examples, high-quality genetic maps or comparative approaches utilizing synteny between genomes.

Accompanied by on-going improvement in DNA sequencing technologies, WGS approaches considerably accelerated plant genome research. The first WGS-based genome sequences were generated for, e.g., black cottonwood *(189)* and grapevine *(190)* (Fig. 1.6). Furthermore, hybrid approaches, which combined classical Sanger sequencing and high-throughput NGS methods as well as BAC-by-BAC and WGS strategies, were also successfully applied to generate draft genome assemblies including, for example, the barley genome *(191)*, the potato genome *(192)* and the tomato genome *(193)*. During the time of this thesis, high-depth NGS whole-genome sequencing utilizing different technologies and library constructions allowed generating draft genome assemblies for the diploid Triticeae *Ae. tauschii (43)* and *T. urartu (44)*. Today, more than 50 draft genome sequences have been published *(194)* encompassing a wide range of different species including model plants [e.g. *Brachypodium distachyon (42)*], non-model plants [e.g. *Phyllostachys heterocycla (195)* or *Capsella rubella (196)*], and agroecological im-

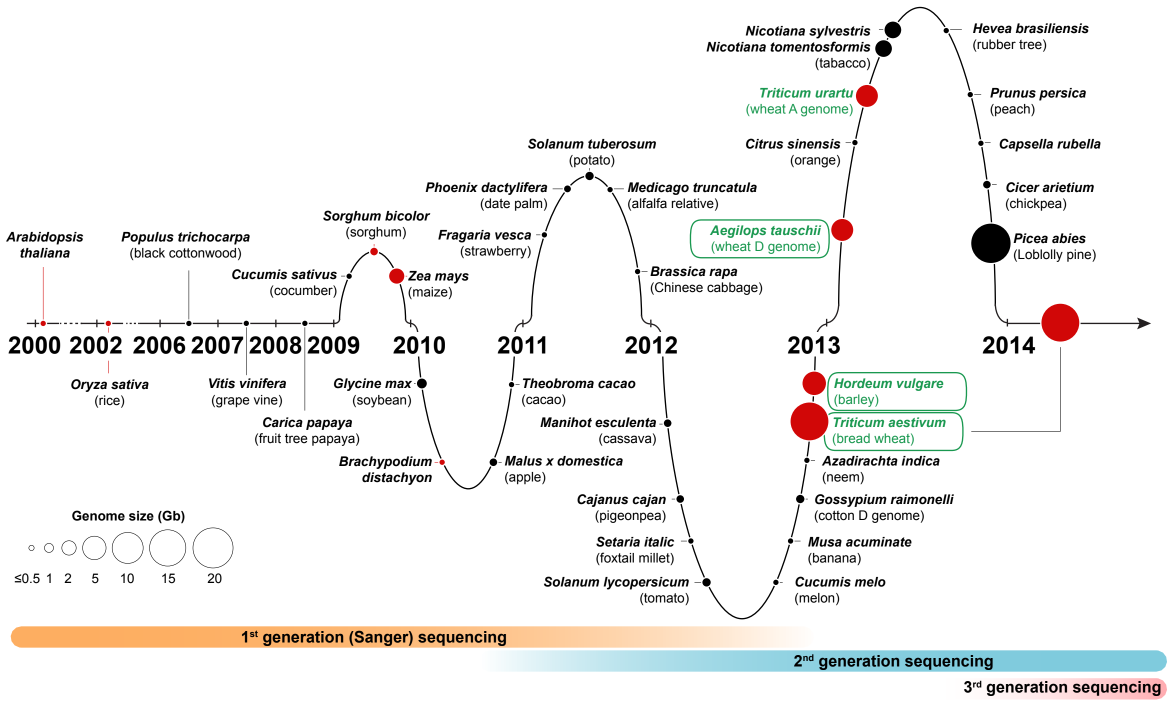


Fig. 1.6. The progress in plant genome sequencing.
Since the release of the *A. thaliana* genome more and more draft or complete plant genome sequences have became available. The progress in plant genome sequencing has been considerably accelerated with second generation sequencing technologies. Future genome sequences will additionally make use of third generation sequencing methods and will profit from significantly larger read length. Green labels mark Triticeae genomes. Dot size corresponds to genome size and red dots depict genome resources utilized in this thesis. Rectangular borders highlight my personal contributions to the genome sequencing projects. [*This overview is a summary of selected genomes based on listings in (22,194).*]

portant crop genomes [e.g. *Oryza sativa (45)*, *Zea mays (41)*, *Gossypium raimondii (88)* or *Solanum tuberosum (192)*].

1.5.2 Challenges and approaches for the analysis of Triticeae genomes

Despite successful application of different first- and second generation sequencing technologies, the high genome complexity still constitutes major bottlenecks for grass genome research. On contrary to many other plant genomes that have been sequenced so far, grass genomes are considerable larger, usually up to several gigabases (Gb) for diploid Triticeae genomes like barley (approximately 5 Gb) and rye (approximately 8 Gb) *(49,51,197)* (Fig. 1.6). Moreover, different degrees of polyploidy inflate grass genome sizes to approximately 10 Gb for allotetraploid *T. turgidum* (pasta wheat) or 17 Gb for allohexaploid *T. aestivum* (bread wheat). In particular, the assembly of ancient or innate polyploid genomes is essentially aggravated by high sequence similarity between duplicated regions that are retained in the (partially) diploidized genome or between homoeologous chromosomes. For example, coding regions of the A, B and D genomes of bread wheat differ in less than 3% of nucleotide positions *(123)*, which complicates determining the genome-of-origin for the obtained short NGS reads. Therefore, homoeologous sequence reads obtained with WGS approaches might be collapsed during sequence assembly *(198)*. Moreover, grass and Triticeae genomes are dominated by repetitive sequences *(51)*. Up to 80% of Triticeae genomes are related to transposable elements (TEs) *(199)*, a significant higher proportion compared to other genomes like *Arabidopsis* with less than 10% TE genome sequences *(137)*. Limitations in sequencing large DNA fragments spanning repetitive sequences impede the complete reconstruction of a contiguous genome sequences *(171,172)*. Therefore, current genome sequences are often restricted to gene containing and low-copy regions, although increased sequencing depth and usage of long-distance mate pair sequencing libraries improve the length of assembled contigs and scaffold *(41,191,192)*.

Large public sequence repositories were established aiming at supporting Triticeae genome *(200,201)* and transcriptome *(202)* analysis through EST *(203)* and full length (fl)-cDNA *(204)* collections. Furthermore, isolation of single chromosomes with flow cytometry sorting technology reduces the complexity of an whole genome approaches to individual chromosome arms, allowing to construct chromosome-specific BAC or NGS libraries *(205–208)*. This approach proves particularly valuable to separate the A, B and D genomes of bread wheat prior to sequencing *(209,210)* and facilitates to assemble each chromosome separately without risking to collapse homoeologous copies *(198)*.

Furthermore, the emergence of high-quality genomes of the closely related species *Brachypodium distachyon (42)*, *Oryza sativa (45)* and *Sorghum bicolor (40)*, allowed implementation of bioinformatic approaches investigating and comparing the structure of Triticeae genomes. One approach, termed the ”GenomeZipper”, exploits large-scale syntenic conservation of gene order among grass genomes (Section 1.2.3) and combines the known gene orders in *Brachypodium*, rice and sorghum with genetic maps in order to approximate a linear positioning along chromosomes. This approach has been successfully applied for the barley genome *(61,62)*, single wheat chromosomes *(63)*, rye *(72)* and perennial ryegrass *(36)* and provided accurate structural information beneficial, for example, construction of physical maps *(211)*, accelerated development of genetic marker maps *(60)* or identification of quantitative trait loci *(212)*.

1.6 Research questions and objectives of this thesis

This dissertation focussed on a comprehensive computational-based analysis of Triticeae genomes. In particular, the implementation and application of bioinformatic approaches to investigate the genome and transcriptome of allohexaploid *T. aestivum* (bread wheat), one of the agroecologically most important cereals, will be discussed in the following chapters. The central question for all conducted experiments was the impact of polyploidization on the genome structure, content and evolution and on the inter- and intra-genomic regulatory interactions orchestrating gene expression among the homoeologous genomes. So far, those genome-wide studies have been limited by the size and complexity of the bread wheat genome. However, due to its economic and industrial importance, the upcoming demographic and environmental challenges and concerns of global food security, significant international efforts have been undertaken to establish comprehensive genome and transcriptome sequence resources for wheat research. Classical and NGS-based DNA sequencing technologies as well as BAC-by-BAC, chromosome-sorting and WGS strategies have been applied and require bioinformatic approaches integrating these heterogenous data resources to bridge structural, evolutionary and functional aspects and, consequently, to contribute to a genome-wide understanding of Triticeae genomes.

Technical aspects of this thesis

Investigation of the biological mechanisms underlying evolution, regulation and traits of bread wheat requires comprehensive genome resources. Therefore, the central technical aspect of this thesis aimed at the development and application of computational workflows to establish sequence catalogues suitable for the genome-wide analysis of polyploid genomes. Novel bioinformatic strategies and integrative concepts were necessary to combine first- and second generation sequencing data with available genome information from closely related reference species. Technical challenges, including fragmentation of genes on multiple contigs during the assembly process, and biological challenges, including highly similar homoeologous sequences, large proportion of deteriorated (pseudo-)genes, asked for adequate analytical strategies. Approaches utilizing comparative-genomics or flow cytometry-sorting to separate chromosome arms constituted major promising starting points to unlock the wheat gene catalogue with homoeolog-specific resolution.

Biological research questions of this thesis

Based on these resources, this work aimed at contributing to an understanding of the genome architecture and regulatory mechanisms for bread wheat. Various studies have shown considerable effects of polyploidization on the genome sequence of ancestral, innate and synthetic polyploids, however, the genome-wide extent of genomic alterations has remained an open question. Quantification of gene loss, retention or duplication rate and subsequent comparative analysis between the hexaploid wheat gene repertoire with that of related diploid genome will give detailed insights in the evolutionary fate of homoeologous genes during polyploid progeny.

Furthermore, genome asymmetry and homoeolog-specific gene expression patterns have been observed for selected genes or gene families, but at the whole genome level the extent and patterns of gene expression divergence between genomes in different tissues has been largely unknown. Global analysis were required to answer, whether polyploidy impacts transcriptional regulation in a sporadic mode, is orchestrated among genomes or affects systematically certain pathways or cellular functions. High-throughput RNA-seq of the bread wheat transcriptome during grain development will add functional insights in order to exploit differences among homoeologous gene expression patterns.

1.7 Overview of this thesis

As outlined in the following, this thesis is divided into three, mostly self-contained chapters (**Chapters 3 to 5**). These constitute autonomous studies with particular experimental designs and focus on distinct research questions. Chronologically ordered, the chapters aim at providing fundamental insights in the genomic landscape of the bread wheat genome and transcriptome, present different but complementary approaches and cross-referencing each other. Thereby, especially chapter 5 relies on genomic resources generated within the predecessing chapter.

To begin with, **Chapter 2**, will introduce the genomic and transcriptomic data sets used in this thesis. The different resources are briefly described and their main usage linked to the individual chapters and underlying research questions.

Chapter 3 will then present the implementation, evaluation and application of a novel gene-centric assembly strategy for the analysis of complex and polyploid genomes based on a comparative genomics and whole genome shotgun sequencing. Applied on the bread wheat genome, this approach allowed assembling a large proportion of the protein-coding genome space without collapsing homoeologous sequences. Quantification of gene retention, gain and losses in hexaploid bread wheat combined with estimates for diploid *Ae. tauschii*, the diploid progenitor of the wheat D genome, and orthologous gene family sizes in fully sequenced and annotated reference plant genomes, will reveal genome dynamics of polyploid evolution. Furthermore, for a substantial number of genes this chapter will show that the OGA provides a suitable framework and will discuss signatures of pseudogene formation in the grasses.

So far, high sequence similarity between homoeologous sequence copies and large stretches of repetitive DNA have precluded the generation of a bread wheat reference genome sequence assembly. To overcome this challenge the International Wheat Genome Sequencing Consortium (IWGSC) applied a ”divide and conquer” approach and isolated DNA of individual chromosome arms by using chromosome sorting technology, which then were separately shotgun sequenced and *de novo* assembled. **Chapter 4** will present the implementation of an extrinsic gene prediction pipeline for the annotation of the ”chromosomal survey sequence” (CSS) assembly. A comprehensive gene set providing sequences and structures for more than 90% of the bread wheat genome was generated allowing to elucidate the structural characteristics of the identified wheat genes and to investigate the presence of thousands of putative non-coding but transcriptional active genomic regions. Targeted gene family analysis will deepen the understanding of the composition of wheat gene families on a chromosome (arm) level. This chapter will also make use of RNA-seq data to analyse and discuss the alternative splicing landscape in bread wheat.

Chapter 5 will show how to make use of the established wheat reference genome assembly and gene annotation for gene expression analysis. By applying high-throughput transcriptome sequencing the spatio-temporal interplay of gene expression regulation in the major cell types of developing wheat endosperm was investigated for three important time points. This chapter will reveal and discuss divergence in gene expression of homoeologous genes, genome asymmetry and biased contribution of individual wheat genomes to particular cellular functions. By using a comparative projection of the wheat genes along the ancestral gene order of seven Triticeae prototype chromosomes, moreover, the impact of chromosomal position on gene expression and formation of chromosomal domains was elucidated. On several layers potential genetic and epigenetic regulatory mechanisms that partially orchestrate inter- and intragenomic gene expression in allohexaploid wheat will be addressed.

Finally, **Chapter 6** will summarize the scientific achievements and discuss possible extensions of this work as well as potential future projects.

Chapter 2

Materials – the utilized genome and transcriptome resources

This thesis encompassed efforts of three international collaborations aiming at a comprehensive characterization of the genome and transcriptome of bread wheat, the economically most important Triticeae genome (Fig. 2.1). Two projects, one conducted together with a research team lead by Prof. Dr. Neil Hall [University of Liverpool, Liverpool, United Kingdom (UK)] and Prof. Dr. Michael Bevan (John Innes Centre, Norwich, UK) (Chapter 3) as well as a second in frame of the International Wheat Genome Sequencing Consortium (Chapter 4), focussed on the generation of genomic resources for bread wheat and the subsequent application to investigate genome dynamics following polyploidization. The third project, initiated by researchers from the

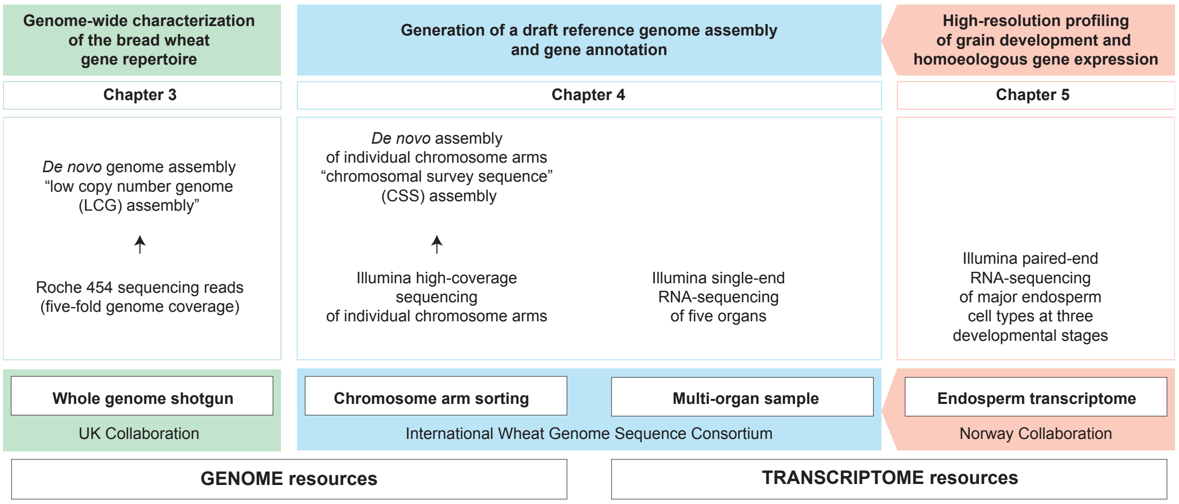

Genome-wide characterization of the bread wheat gene repertoire
Generation of a draft reference genome assembly and gene annotation
High-resolution profiling of grain development and homoeologous gene expression
Chapter 3
Chapter 4
Chapter 5
De novo genome assembly "low copy number genome (LCG) assembly"
Roche 454 sequencing reads (five-fold genome coverage)
De novo assembly of individual chromosome arms "chromosomal survey sequence" (CSS) assembly
Illumina high-coverage sequencing of individual chromosome arms
Illumina single-end RNA-sequencing of five organs
Illumina paired-end RNA-sequencing of major endosperm cell types at three developmental stages
Whole genome shotgun
Chromosome arm sorting
Multi-organ sample
Endosperm transcriptome
UK Collaboration
International Wheat Genome Sequence Consortium
Norway Collaboration
GENOME resources
TRANSCRIPTOME resources

Fig. 2.1. Datasets analysed in this thesis with respect to projects and biological questions. This thesis combines genome and transcriptome resources generated within three different projects and consortia utilizing different next generation sequencing technologies.

Norwegian University of Life Sciences led by Prof. Dr. Odd-Arne Olsen, made use of the established resources to study the transcriptional landscape of allohexaploid wheat and conducted cell type-specific profiling of gene expression for developing wheat endosperm (Chapter 5). Each consortia developed and applied different experimental and analytical concepts for the complementary study of particular genomic and transcriptomic questions. As briefly summarized in the following of this chapter, all projects utilized different next generation sequencing technologies to generate highly heterogeneous data sets, which required different bioinformatic processing and analysis.

2.1 Whole genome shotgun sequencing of the bread wheat genome (UK collaboration)

Within the UK research collaboration a whole genome shotgun data set of bread wheat cultivar ”Chinese Spring” *(213)*, the best studied wheat cultivar *(198,214)*, was generated at the John Innes Centre (Norwich, UK) by using 454 pyrosequencing technology *(154,157)* [GS FLX Titanium and GS FLX1 platforms (Roche Applied Science, Basel, Switzerland)]. The achieved collection of shotgun sequencing reads encompassed a total of 85 Gb of sequence data and 220 mio reads corresponding to approximately five-fold genome coverage (Table 2.1).

Furthermore, the project collaborators at the Centre for Genome Research of the University of Liverpool (UK) filtered the obtained genomic shotgun reads for repetitive sequences and computed a *de novo* genome assembly by using the `gsAssembler`-tool from the Newbler package, an overlap-graph assembly toolbox developed specially for Roche 454 sequencing projects *(154)*. Due to low assembly stringency, i.e. 90% minimum alignment identity for overlapping reads, large proportion of homoeologous (protein-coding) sequence copies were expected to be collapsed (Section 3.1.3). Thus, this assembly was termed ”low-copy-number genome” (LCG) assembly. All sequence resources have been made publicly available with study accession PRJEB217 in the European Nucleotide Archive (ENA) hosted by the European Bioinformatics Institut of the European Molecular Biology Laboratory (EMBL-EBI).

Table 2.1. Sequence statistics of the bread wheat whole genome shotgun data set generated within the UK collaboration.

	Raw 454 sequencing reads	LCG assembly
Number of sequences (mio)	220	5
Total sequence (bp)	82,801,349,875	3,800,325,216
Minimum sequence length (bp)	18	100
Maximum sequence length (bp)	2,032	21,721
Average sequence length (bp)	389	714

Additionally, this work made use of a WGS resource obtained for *Ae. tauschii*, the diploid progenitor of the wheat D genome, in a study of Luo *et al. (215)* (Section 1.2.4). The authors also utilized Roche 454 pyrosequencing technology *(154,157)* to generate genomic sequencing reads, which encompassed a total of 12.8 Gb sequence and represented approximately threefold genome coverage. This data set has been made publicly available in the Sequence Read Archive (SRA) under the study accession SRP012566 and was retrieved from there.

2.2 Resources for the gene annotation of chromosomal sequence assemblies of the bread wheat genome (IWGSC consortium)

Aneuploid bread wheat lines derived from double ditelosomic stocks of the hexaploid wheat cultivar ”Chinese Spring” *(213)* were used to isolate and purify DNA of individual chromosome arms by flow-cytometric sorting at the Centre of Plant Structural and Functional Genomics (Olomouc, Czech republic) *(205–208)*. Except for 3B, which could be isolated as a complete chromosome, individual chromosome arms were sequenced to a depth between 30-fold and 241-fold with Illumina sequencing instruments *(155,156)* [HiSeq 2000 or Genome Analyser IIx (Illumina Inc., San Diego, California, USA)] to generate 100 or 150 base paired-end reads (Table 2.2). The obtained reads were *de novo* assembled for each individual chromosome arm with the short-read *de novo* assembler ABySS *(164)*. The generated chromosomal sequence survey assemblies were checked for contaminations and, if necessary, cleaned and re-assembled. Repetitive sequences mainly assembled into small contigs with less than 200 bp length and were excluded from the final assembly of 10.2 Gb (10.5 mio contigs). Sequencing and assembly has been carried out by collaborators at The Genome Analysis Centre (Norwich, UK) and resources have been made publicly available in the ENA (study accession PRJEB3955).

Repetitive sequences were masked for individual chromosomes based on sequence homology searches against the MIPS-REdat Poaceae library[(1)], which includes repetitive sequences from public available plant repeat databases and from *de novo* detection of long terminal repearetrotransposons in grass genomes. Matching sequences against the repeat catalogue were masked by ”N”s and contigs with stretches of less than 100 bp unmasked sequences removed. This strategy resulted in a final repeat-masked version of the CSS assembly including 1.7 mio contigs and with a L50 of 5,858 bp.

For gene annotation on basis of the CSS assembly a multi-organ RNA-seq collection was prepared including five tissues (root, leaf, grain, stem and spike) of bread wheat cultivar ”Chinese Spring” each sampled at three developmental stages. RNA from the same organ was pooled and each library was sequenced to 101 base single-end reads on the Illumina HiSeq 2000 sequencing machines *(155,156)* (Illumina Inc., San Diego, California, USA) (Table 2.3). This data set was generated by INRA (URGI – Research Unit in Genomics-Info, Versailles, France) and sequencing reads have been made publicly available in the ENA (study accession PRJEB4750).

[(1)] The MIPS-REdat *Poaceae* repeat library was downloaded from http://mips.helmholtz-muenchen.de/plant/recat (version 8.6.2).

Table 2.2. Sequence and assembly statistics of the chromosomal survey sequence assembly generated by the IWGSC.

Chr. arm	Mb[(a)]	x-fold[(b)]	CSS[(c)]				Rep.-masked CSS[(c)]	
			#[(d)]	Σ (Mb)[(d)]	L50 (bp)	GC (%)	#[(d)]	L50 (bp)[(d)]
1AS	275	137	187,49	178.1	2,242	45.8	34,793	4,769
1AL	523	80	197,674	250.0	2,639	44.5	26,746	6,369
2AS	391	120	264,555	255.2	2,398	45.6	34,722	6,678
2AL	508	97	321,517	328.2	2,688	45.3	45,893	6,677
3AS	360	36	242,308	201.8	1,404	45.2	33,943	3,846
3AL	468	30	303,844	247.2	1,346	44.8	43,823	3,789
4AS	317	241	301,954	282.3	2,782	45.6	32,079	7,499
4AL	539	116	362,01	356.0	2,969	45.3	64,364	6,601
5AS	295	67	182,938	198.8	3,509	44.0	19,719	8,713
5AL	532	48	403,265	318.1	2,078	42.9	47,572	5,355
6AS	336	106	210,388	219.2	2,669	45.6	28,041	7,091
6AL	369	92	245,867	214.4	2,154	45.0	34,03	6,589
7AS	407	28	262,653	198.0	1,47	44.1	44,175	4,397
7AL	407	46	233,306	252.4	2,271	44.0	35,586	5,849
Σ A	5,727	89	3,225,219	27,761	2,235	44,8	494,859	6,016
1BS	314	143	181,801	212.8	3,287	46.0	26,05	7,413
1BL	535	63	198,968	299.4	3,12	44.2	29,783	7,151
2BS	422	121	244,668	292.0	3,711	46.5	35,743	8,069
2BL	506	142	365,563	404.5	2,941	45.5	75,879	6,89
3B	993	89	546,922	638.6	2,655	46.0	75,022	6,855
4BS	319	196	274,504	308.2	3,463	46.4	38,515	8,755
4BL	430	55	317,294	248.7	1,974	45.0	46,576	5,883
5BS	290	146	137,38	174.5	3,315	46.9	18,001	7,365
5BL	580	107	436,173	415.2	2,924	46.5	75,887	7,537
6BS	415	72	166,632	210.2	2,366	42.9	29,566	4,972
6BL	498	105	251,706	257.4	2,031	44.4	35,727	4,824
7BS	360	59	178,789	206.1	2,428	43.7	24,119	6,435
7BL	540	37	328,725	259.6	1,556	43.5	58,554	4,144
Σ B	6,202	103	3,505,483	36,644	2,536	45,2	545,977	6,161
1DS	224	157	126,156	128.2	2,85	46.5	17,725	6,622
1DL	381	125	292,785	254.4	2,561	43.7	35,77	6,297
2DS	316	147	245,107	166.0	1,241	44.6	43,044	4,635
2DL	411	113	508,239	261.6	701	44.1	110,446	3,247
3DS	321	85	314,944	145.0	515	42.1	46,795	1,697
3DL	449	32	326,758	186.0	967	42.1	69,259	2941
4DS	231	173	118,29	142.1	3,278	46.6	18,245	7428
4DL	416	154	454,216	254.4	816	45.7	197,398	1855
5DS	258	220	148,048	148.0	2,353	46.8	22449	5945
5DL	490	94	223,456	236.8	2,647	44.6	34622	7049
6DS	324	68	88,542	156.6	4,297	44.9	16,077	8,904
6DL	389	76	203,805	199.8	2,077	45.3	26,236	6,821
7DS	381	54	216,406	209.1	1,967	44.2	36,701	5,031
7DL	346	77	161,061	222.9	3,638	45.4	26,737	7,399
Σ D	4,937	113	3,321,352	21,304	1,953	44,8	669,311	5,419
Σ	16,866	101	10,052,054	85,709	2,234	45,2	1,710,147	5,858

[(a)] Estimated chromosome arm sizes were taken from *(210)*.

[(b)] Sequencing read depth for individual chromosome arms (x-fold).

[(c)] Statistics for original and repeat-masked version of the CSS assembly.

[(d)] Number (#) and total (Σ) sequence of assembled contigs $\geq$200 bp.

Table 2.3. Sequence statistics of the multi-organ RNA-seq generated for gene annotation by the IWGSC.

Tissue	ID	Read length	Reads (mio)	Sequence
grain	GRA	101 bp	117.7	11.9 Gb
leaf	LEA	101 bp	127.0	12.8 Gb
root	ROO	101 bp	112.5	11.4 Gb
spike	SPI	101 bp	140.0	14.1 Gb
stem	STE	101 bp	118.1	11.9 Gb
Σ	-	-	615.3	62.1 Gb

2.3 Analysis of the transcriptome in developing bread wheat endosperm (Norway collaboration)

Starch and storage proteins constitute the major ingredient of most cereal end-products. As these compounds are mainly accumulated in the nuclear endosperm of flowering plants, including the cereals maize, rice, barley and wheat, understanding the grain transcriptome is of large industrial relevance. The cereal endosperm consists of three major cell types, aleurone (AL), starchy endosperm (SE) and transfer cells (TC), which have spatially and temporally distinct morphological structures and adopt different functional responsibilities *(216,217)* (Chapter 5). To investigate gene expression in this important organ, collaborators of the Norwegian University of Life Sciences (Ås, Norway) constructed RNA-seq libraries for distinct cell types of developing endosperm (Table 2.4). Therefore, wheat plants were grown in two greenhouses (GH) and grains harvested at 10, 20, and 30 days post anthesis (DPA). These were further manually dissected into aleurone, transfer cells and starchy endosperm under the dissection microscope. Thereby, to exclude any later bias in the analysis by using the IWGSC genome sequence assembly as reference, seeds

Table 2.4. Sequence statistics of the wheat endosperm transcriptome data set generated within the Norway collaboration.

Sample	GH	BR	Read pairs	Reads	Sequence (bp)	Σ read pairs	Σ sequence (Gb)
10 DPA W	1	1	20,361,333	40,722,666	4,112,989,266		
	1	2	26,791,465	53,582,930	5,411,875,930		
	2	1	30,235,123	60,470,246	6,107,494,846		
	2	2	33,413,758	66,827,516	6,749,579,116	110,801,679	22,38
20 DPA W	1	1	34,617,242	69,234,484	6,992,682,884		
	1	2	30,517,594	61,035,188	6,164,553,988		
	2	1	28,011,277	56,022,554	5,658,277,954		
	2	2	32,249,714	64,499,428	6,514,442,228	125,395,827	25,33
20 DPA AL	1	1	32,919,785	65,839,570	6,649,796,570		
	1	2	30,833,988	61,667,976	6,228,465,576		
	2	1	27,753,881	55,507,762	5,606,283,962		
	2	2	31,365,012	62,730,024	6,335,732,424	122,872,666	24,82
20 DPA SE	1	1	30,009,734	60,019,468	6,061,966,268		
	1	2	29,714,230	59,428,460	6,002,274,460		
	2	1	26,664,432	53,328,864	5,386,215,264		
	2	2	27,602,634	55,205,268	5,575,732,068	113,991,030	23,03
20 DPA TC	1	1	18,586,985	37,173,970	3,754,570,970		
	1	2	31,121,623	62,243,246	6,286,567,846		
	2	1	29,885,904	59,771,808	6,036,952,608		
	2	2	29,668,161	59,336,322	5,992,968,522	109,262,673	22,07
30 DPA ALSE	1	1	31,433,795	62,867,590	6,349,626,590		
	1	2	22,422,406	44,844,812	4,529,326,012		
	2	1	29,554,700	59,109,400	5,970,049,400		
	2	2	29,381,216	58,762,432	5,935,005,632	112,792,117	22,78
30 DPA SE	1	1	23,711,650	47,423,300	4,789,753,300		
	1	2	27,182,660	54,365,320	5,490,897,320		
	2	1	37,524,396	75,048,792	7,579,927,992		
	2	2	25,114,866	50,229,732	5,073,202,932	113,533,572	22,93
Σ						808,649,546	163,35
20 DPA AL	1	1*	32,374,902	64,749,804	6,539,730,204		
20 DPA AL	1	1*	32,685,090	65,370,180	6,602,388,180		

Numbering indicates greenhouses (GH) and biological replicates (BR). Stars mark technical replicates.

from the same variant of *T. aestivum* cultivar ”Chinese Spring” that was used for generating the reference genome sequence, were provided by Bikram Gill (Kansas State University, Manhattan, Kansas, USA).

A total of 30 mRNA samples were prepared and sequenced including two biological replicates (BR) for seven samples of plants grown in two greenhouses (2 BR $\times$ 2 GH $\times$ 7 conditions = 28 libraries) as well as two additional technical replicates for one sample by using paired-end HiSeq2000 technology *(155,156)* (Illumina Inc., San Diego, California, USA) with an average insert size 200 bp (Table 2.4). The high-throughput sequencing yielded in 110 mio (20 DPA TC) to 125 mio (20 DPA AL) read-pairs per endosperm sample and in a total of 809 mio read-pairs (163 Gb raw sequence). Sequencing data has been made publicly available in the ArrayExpress database hosted by the EBI (accession E-MTAB-2137).

Chapter 3

Genome dynamics of polyploid bread wheat

Whole genome shotgun sequencing is a rapid, cost and time efficient way to generate large genomic resources by sequencing of randomly-fragmented DNA clones *(218)* (Section 1.4.1). However, the assembly and computational analysis of obtained WGS sequence reads is substantially complicated for most plants because of the large genome sizes and high genome plasticity due to repetitive sequences *(51,219)* and different degrees of polyploidy *(220)* (Section 1.4.2). This applies especially to allohexaploid bread wheat (*T. aestivum* L.), which is one of the largest plant genomes arising by reason of two hybridization events that brought together three diploid genomes (2n=6x=42; AABBDD) *(66,67)* (Section 1.2.4). The sequences of these three homoeologous genomes were found to be highly similar among each other *(123)*. Consequently, distinguishing the genome-of-origin for individual reads in the pool of WGS data is substantially hampered, if not impossible.

To overcome this challenge a novel comparative genomics-based assembly concept, the ”orthologous group assembly” (OGA), was developed in this thesis. The major goal of the OGA was to generate homoeologous-specific sequence assemblies based on WGS sequence data for highly complex and polyploid genomes. In contrast to traditional *de novo* genome assembly and analysis concepts, the OGA focused primarily on the protein-coding portion of a genome. Therefore, available protein sequences of closely related grass genomes were used to define an orthologous gene family framework restricting the search space onto genes that are conserved among related taxa. The obtained bread wheat WGS sequencing reads were projected onto orthologous protein sequences and, separately for each protein, assembled applying highly stringent criteria. This approach minimized collapsing homoeologous sequence copies and allowed further quantification of distinct gene copies for the bread wheat genome.

On the one hand, this chapter will describe the underlying technical concepts of the OGA including the definition of an orthologous gene family framework for bread wheat, the computational estimation of gene copy numbers and the evaluation of the implemented strategy using *in silico* simulation experiments. On the other hand, gene content dynamics following polyploidization will be investigated by comparing the gene family sizes in the hexaploid wheat genome with that in the diploid D-genome progenitor *Ae. tauschii* (2n=2x=14; DD). Furthermore, this chapter will discuss the extent and the potential influence of pseudogene formation on the genome structure and the evolution of gene families of one of the world's most important crops.

All results shown in this chapter are part of following publications:

- **Analysis of the bread wheat genome using whole genome shotgun sequencing**
R. Brenchley‡, M. Spannagl‡, **M. Pfeifer‡**, G. L. A. Barker‡, R. D'Amore‡, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou, D. Bolser, S. Kay, D. Waite, M. Trick, I. Bancroft, Y. Gu, N. Huo, M. C. Luo, S. Sehgal, B. Gill, S. Kianian, O. Anderson, P. Kersey, J. Dvorak, W. R. McCombie, A. Hall, K. F. X. Mayer, K. J. Edwards, M. W. Bevan and N. Hall
Nature. 491(7426):705–710, 2012.
‡ joint first authors

- **Analysing complex Triticeae genomes – concepts and strategies**
M. Spannagl, M. M. Martis, **M. Pfeifer**, T. Nussbaumer and K. F. X. Mayer
Plant Methods. 6;9(1):35, 2013.

3.1 Homoeologous-specific sequence analysis of the bread wheat genome

All methods described in this chapter were specifically developed for the analysis of the bread wheat genome. However, the underlying concept can be transferred to any other complex or polyploid genome. The following experiments were based on whole genome shotgun sequences obtained for the bread wheat genome with approximately five-fold genome coverage by using Roche 454 pyrosequencing technology (Section 2.1). Importantly, the obtained reads (average read length of 388 bp) were expected to be of sufficient length to distinguish homoeologous (protein-coding) sequences based on genome-specific SNPs, which have been reported to occur with a frequency of one per 145 base pairs reported for coding sequences of homoeologous genes *(123)*.

3.1.1 Definition of an orthologous gene family framework for the grasses

The orthologous group assembly aimed at reducing the analysis complexity by focussing towards the protein-coding sequences of the genome. Therefore, the construction of an orthologous gene family framework, which represents comprehensively the wheat gene space, was a key requirement of the OGA that enabled screening for gene candidates in the WGS data[1]. For this purpose the reference genomes of *Brachypodium* (*Brachypodium distachyon*) *(42)*, rice (*Oryza sativa*) *(45)* and sorghum (*Sorghum bicolor*) *(40)* as well as a collection of more than 23,000 public available barley (*Hordeum vulgare*) fl-cDNAs *(221)* provided particularly valuable protein sequence information to reconstruct conserved gene families from different grass sub-families spanning an evolutionary time frame of approximately 45 mio to 60 mio years *(42)* (Fig. 1.2). A total of 86,944 sequences, derived from protein-coding genes of these three grass genomes and peptide predictions *(222)* of the barley fl-cDNAs *(221)*, were clustered into 20,496 orthologous groups of putative orthologous genes and close paralogs by using the OrthoMCL software *(223)* (version 1.4). These groups were defined by proteins of at least two species, thus, represented a set of well-conserved gene families among the grasses. Almost all orthologous groups [20,051 (98%)] were detected by stringent peptide sequence comparison to the LCG assembly of the bread wheat genome utilizing the "Basic Local Alignment Search Tool" (BLAST) with a maximum Expect (*E*) value of 10^{-10} and the BLASTX option. This assembly was provided by the UK collaboration partners, who filtered repetitive shotgun sequencing reads and assembled the remaining genomic sequences *de novo* using a classical OLC-based assembly approach with relaxed overlap thresholds (Section 2.1). For each orthologous group the reference protein that was best represented in the wheat sequences was defined as orthologous group representative (OGR) serving as a template protein in the OGA and the subsequent analysis (Table 3.1).

The protein sequences of the selected OGRs were further compared against metabolic genes in *A. thaliana (137)* (90% detected), publicly available wheat fl-cDNAs *(204)* (92% detected) and cDNA assemblies from the wheat HarvEST database *(203)* (version 1.19 stringent) (78% detected). The high level of captured genes participating in major plant pathways, good coverage of wheat cDNA sequences and high detection rate by the wheat LCG assembly suggested that the selected OGRs provided a suitable framework for further comprehensive analysis.

Table 3.1. Number of orthologous groups defined in the gene family framework built on basis of high-quality protein sequences of related grasses.

	Number of groups	**Alignment identity[a]**
Total orthologous groups clusters	20,496	-
Total orthologous groups with OGR	20,051	-
Brachypodium	7,996	75%
Barley fl-cDNAs	5,337	80%
Rice	3,136	70%
Sorghum	3,582	70%
Total orthologous groups without OGR (no homology support in the wheat LCG assembly)	445	-

[a] Minimum alignment identity thresholds for alignments of genomic wheat sequencing reads.

[1] I gratefully acknowledge Manuel Spannagl for the definition of the orthologous gene family framework and his close collaboration in the selection of the orthologous group representatives and in the evaluation of the comprehensiveness of the defined framework.

3.1.2 The orthologous group assembly and calculation of the wheat gene copy number

The analytical workflow of the orthologous group assembly included three consecutive steps: (i) pre-processing of the raw sequencing reads, (ii) identification of genic wheat sequences and their allocation to OGRs and (iii) stringent assembly of the assigned genomic shotgun sequence reads individually for each OGR (Fig. 3.1a). Based on coverage and alignment depth of the orthologous group representatives by the consensus assemblies ("sub-assemblies"), the gene copy number in bread wheat (Fig. 3.1b) was predicted, further enabling to monitor and quantify genome dynamics in the polyploid genome.

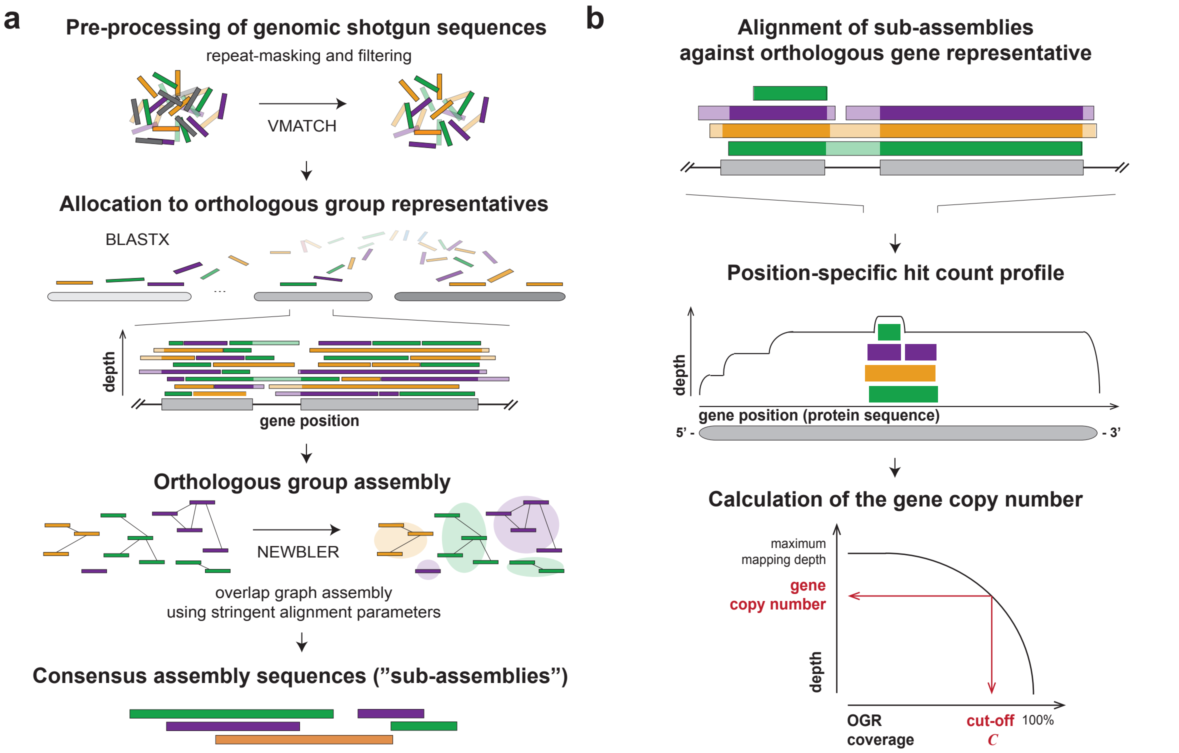

a
Pre-processing of genomic shotgun sequences
repeat-masking and filtering
VMATCH
Allocation to orthologous group representatives
BLASTX
...
depth
gene position
Orthologous group assembly
NEWBLER
overlap graph assembly
using stringent alignment parameters
Consensus assembly sequences ("sub-assemblies")
b
Alignment of sub-assemblies
against orthologous gene representative
Position-specific hit count profile
depth
gene position (protein sequence)
5' -
- 3'
Calculation of the gene copy number
maximum
mapping depth
gene
copy number
depth
OGR
coverage
cut-off
C
100%

Fig. 3.1. The orthologous group assembly and the estimation of gene copy number.
a, Genomic shotgun reads were repeat masked and assigned to corresponding orthologous gene representatives. Each sequence bin is separately assembled and consensus ("sub-assembly") sequences generated based on overlaps between reads. **b,** The sub-assembly sequences were aligned to the cognate OGR and ordered along its protein sequence. Then, the alignments were transferred into a position-specific hit count profile that counts the number of distinct sub-assemblies mapped to each amino acid of the template protein. The wheat gene copy number was computed as the maximum number of distinct sub-assemblies covering a defined proportion of the protein-coding sequence of the OGR, which was defined by a coverage cut-off C. Grey boxes represent the protein sequence of orthologous group representative, whereas lines connecting boxes depict exon boundaries. Coloured boxes visualize sequencing reads and assembled sequences, respectively. The colour code groups sequences that originate from the same genome and light colouring visualize non-coding regions.

Pre-processing of genomic shotgun reads

Repetitive sequences represent the largest fraction of DNA sequences in grass genomes *(51,219)* and considerably extend search space and computational complexity. Since this study was primarily focussed on the protein-coding portion of the wheat genome, reads related to known repeat sequences were filtered prior to the search phase of the OGA. Besides an improved computational efficiency of the implemented assembly protocol by decrease in memory and time requirements as well as simpler data handling and processing (e.g. homology search against the OGRs), the removal of repetitive sequences also avoided overestimation of the computed wheat gene copy numbers. Repetitive mechanisms and transposable element activity have been reported to capture, integrate and amplify gene fragments and were frequently associated with the generation of pseudogenes *(199,224,225)*. Thus, repetitive sequences would substantially effect downstream analysis and might inflate gene family sizes.

To identify repetitive sequences, I compared the entire collection of sequence reads against the MIPS-REdat *Poaceae* repeat library[2] by using VMATCH *(226)* with default parameters and minimum 70% sequence identity over at least 100 bp length (parameters: -identity 70 -l 100). Matching sequences were masked by ”N”s. Additionally, reads without stretches of at least 50 bp unmasked nucleotides were removed and excluded from further analysis.

Overall, a total of 62.3 Gb out of 82.8 Gb (75%) raw genomic sequence showed significant homology to known repeat elements (Fig. 3.2a). This was largely consistent with an estimated repeat content of about 70% for bread wheat *(51,214,219)*. Cleaning highly repetitive sequencing reads reduced the search space by more than two thirds and passed 65.8 mio reads [24 Gb out of 83 Gb (29%)] to the subsequent step in the OGA (Fig. 3.2b).

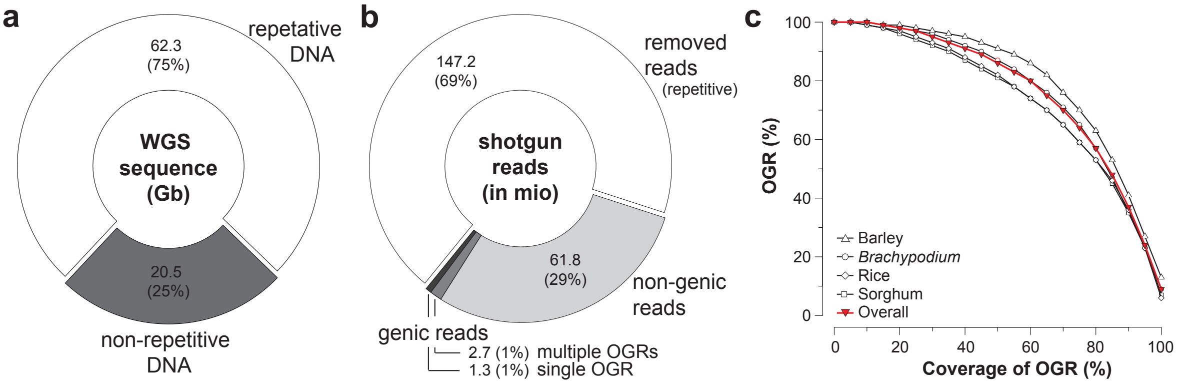


Fig. 3.2. Repeat-masking, filtering and mapping statistics of genomic shotgun sequence reads. **a,** Fraction of raw genomic sequence data identified as repetitive. **b,** Shotgun sequences that are entirely composed of repetitive DNA were removed, whereas the remaining reads were aligned against the OGRs. **c,** Cumulative coverage distribution of OGRs by aligned genomic shotgun reads.

[2] The MIPS-REdat *Poaceae* repeat library was downloaded from http://mips.helmholtz-muenchen.de/plant/recat (version 8.6.2).

Allocation of non-repetitive shotgun reads to orthologous gene groups
Next, the remaining shotgun reads were allocated to OGRs based on protein sequence homology deduced from stringent BLASTX alignments ($E \leq 10^{-10}$), which were filtered for alignment length ($\geq$30 amino acids) and sequence similarity. Therefore, I applied different identity thresholds to account for different evolutionary distances between bread wheat and the reference plant genomes used for the definition of the respective OGR *(42)* (Table 3.1). In case of valid alignments of a read to multiple OGRs, the wheat sequence read was assigned to the OGR with the highest-scoring BLASTX *(227)* alignment.

In total, 4 mio shotgun sequences (2%) were aligned and allocated to 19,483 OGRs (97%) (Fig. 3.2b). Approximately two-third (68%) of the mapped genomic shotgun reads matched a single representative gene with the specified alignment parameters. Generally, wheat reads covered the protein-coding sequence of OGRs with high coverage and more than two third of the template proteins in full-length with at least 70% coverage (Fig. 3.2c). However, minor variations in the coverage of OGRs from different genomes reflected the evolutionary distances to wheat and the respective reference species. Whilst barley reference proteins were covered best, OGRs selected from rice and sorghum were less represented in the wheat data set. This observation corroborated recent studies showing that the evolution in gene structure is an important mechanism for functional diversification and gene novelty additionally to exchanges of amino acids *(228–230)*. Overall, high detection rate of OGRs and almost full-length coverage of their protein sequences indicated that the chosen OGRs constituted suitable templates for capturing genic wheat sequences from the WGS data set, which allowed further orthologous-guided analysis.

Generation of gene-centric sub-assemblies
Sequence information and quality scores of aligned shotgun reads were extracted from the original sequence library files and individual assemblies were computed for each OGR by using Newbler *(154)*, a *de novo* overlap-graph assembler optimized for the assembly of shotgun reads obtained with Roche 454 pyrosequencing technology (Section 1.4.2). The detection of overlaps among reads is a major and critical step, in particular, for the assembly of polyploid genomes with highly redundant sequences of different parental origin. While too relaxed minimum overlap identity (mi) for accepting overlaps between reads would collapse homoeologous sequence copies, too stringent parametrization would be sensitive to sequencing errors and, consequently, imply overestimation of the gene copy number.

Therefore, I evaluated the impact of different stringency levels on the OGA and performed separate assemblies using 97% mi, 99% mi and 100% mi, respectively. Although the applied minimum overlap alignment identity parameters differed only by three percent, the resulting OGAs were influenced considerably (Table 3.2). Whereas more than three quarters (76%) of reads were assembled into contigs for 97% mi, the number of assembled reads dropped to 51% requiring perfect alignments between overlapping reads (100% mi). On the contrary, the number of genomic sequencing reads remaining singletons almost doubled between 97% mi and 100% mi and the total assembled sequence increased by 1.6-fold (498 Mb to 794 Mb). Already these statistics emphasized the importance of correctly setting up set-up and evaluation of the assembly protocol, which will be further discussed in the following sections.

Table 3.2. Newbler assembly statistics of orthologous group assemblies with different stringency levels.
Three orthologous group assemblies were performed by using different minimum alignment identity (mi) thresholds to accept overlapping genomic shotgun sequencing reads.

	97% mi	**99% mi**	**100% mi**
No. of excluded reads[(a)]	75,440 (2%)	90,254 (2%)	247,330 (6%)
No. of assembled reads	3,038,943 (76%)	2,689,502 (67%)	2,057,928 (51%)
No. of remaining singletons[(b)]	887,615 (22%)	1,222,242 (31%)	1,696,740 (42%)
No. of assembled contigs	205,817	172,039	120,501
No. of sub-assemblies[(c)]	1,093,432	1,394,281	1,817,241
total sequence (bp)	497,965,174	630,756,335	793,978,129
min. / max. length (bp)	52 / 7,415	52 / 7,312	52 / 4,386
mean length (bp)	455.41	452.39	436.91
L50 / L90 (bp)	482 / 323	479 / 326	471 / 322

[(a)] Problematic, too short or repetitive sequencing reads excluded for the assembly by Newbler.
[(b)] Sequencing reads without any significant overlap to any other sequencing read.
[(c)] Combined set of sequencing reads remaining singletons and assembled contigs.

Calculation of wheat gene copy numbers

To determine the gene copy number, the wheat consensus sub-assemblies were aligned against their cognate OGRs [BLASTX *(227)* (E $\leq 10^{-10}$)] (Fig. 3.1b). Therefore, the same alignment parameters were applied as used for mapping the raw sequencing reads (Table 3.1). All consecutive high-scoring segment pairs of the returned alignments were accepted to account for stretches of non-coding sequences in sub-assemblies, which represent introns and connect two or more neighbouring exons of the OGR. For each OGR the alignments were transferred into a position-specific hit count profile by counting the number of aligned sub-assemblies at each amino acid position along the template protein sequence. Then, the profile was converted into a cumulative coverage distribution, ranging from one to the maximum hit-count in the profile by only considering sequence positions that were tagged by at least one sub-assembly. Finally, the wheat gene copy number was defined as the maximum hit count assigned to C percent coverage of the protein-coding sequence of the cognate OGR. For all subsequent statistical analysis data set was constrained to OGRs that were covered in full-length by wheat sequences (C = 70%) to avoid wrong copy number estimates due to the large number of gene fragments and pseudogenes, which have been shown to be highly abundant in the wheat genome *(199)*.

Calculation of the gene retention rate in hexaploid wheat

To globally characterize the degree of retention, gain or loss of genes on basis of the OGA, the gene retention rate (r) was computed as the ratio between the number of predicted gene copies ($c_{\mathrm{predicted}}$) and the respective gene family size in the reference genomes ($c_{\mathrm{reference}}$):

$$r = \frac{c_{\mathrm{predicted}}}{c_{\mathrm{reference}}} \tag{3.1}$$

For example, in the naïve expectation for the gene retention rate of hexaploid wheat, which assumes complete absence of any genome dynamics, a single gene copy would be present in each wheat genome [$c(\mathrm{A}) = c(\mathrm{B}) = c(\mathrm{D}) = 1$] that corresponds to one gene in the diploid reference genomes [$c(2\mathrm{n}) = 1$], i.e.

$$r_{\mathrm{naïve}}(6n) = \frac{c(\mathrm{A}) + c(\mathrm{B}) + c(\mathrm{D})}{c(2n)} = \frac{1+1+1}{1} = \frac{3}{1} = 3$$

In this study, the gene family sizes of the diploid reference genomes were determined from the number of proteins that clustered with the selected OGRs. These were paired with the predicted gene copy numbers. Then, a locally-weighted polynomial regression *(231)* of the median copy number predictions for each reference gene families size was determined by using the `lowess`-function *(232)* implemented in the R package stats[3]. Thereby, only reference gene families with up to 75 copies were considered as beyond genes were likely to constitute repeat sequences resulting from transposable element activity *(199)*. Hence the steepness of the regression fit defined the gene retention over the whole sample size, the gene retention rate was calculated as the mean gradient of the polynomial approximation at each data point.

3.1.3 Gene copy number estimations for different OGAs

By measuring the alignment depth that specifies the number of distinct sequences aligned over the protein-coding regions of an OGR, a median depth of 13 was observed for repeat-masked genomic shotgun sequence reads (Fig. 3.3a). This was largely consistent with the five-fold coverage of the underlying whole genome sequencing experiment and a hexaploid genome constitution. Furthermore, substantial differences in alignment depth and estimated gene copy numbers were also evident for different minimum overlap identity thresholds (Fig. 3.3b). As already indicated by the assembly statistics before (Table 3.2), applying different assembly stringency levels had also considerable impact on the orthologous group assembly. Due to the high similarity of homoeologous sequences among the A, B and D genomes *(123)*, the majority of homoeologous gene copies were collapsed in the OGA with 97% mi. On the contrary, requiring perfect overlaps among shotgun reads resulted in an alignment depth about six. This indicated that distinct homoeologous gene copies were maintained, however, the observed alignment depth for 100% mi exceeded the expectations for the gene count in an hexaploid genome. Most probable this

[3] http://stat.ethz.ch/R-manual/R-patched/library/stats/html/lowess.html

was caused by sequencing errors, which have been estimated to affect up to one percent of nucleotides in typical genomic fragments generated by using Roche 454 pyrosequencing technology *(233)*. These prevented the assembly of conjugating shotgun reads and, consequently, increased artificially alignment depth as well as the estimated gene copy number.

The alignment depth was also measured for the LCG assembly, which was generated with 90% mi, the default Newbler assembly parameter (Section 2.1). The observed constant median alignment depth of one suggested entire collapse of homoeologous sequences (Fig. 3.3a). This demonstrated the inapplicability of traditional assembly approaches for bread wheat and likely any other complex, young polyploid genome.

Taken together, the distribution of alignment depth over the OGR and the observed frequency distribution of gene copy numbers suggested that assemblies with 99% minimum sequence identity between overlapping shotgun reads would result in the most appropriate sequence collection, which accurately discriminate between homoeologous sequences while simultaneously accounting for technical limitations in form of sequencing errors. However, the absence of genome-wide studies in bread wheat, which would provide a reference for calibration of the assembly parameters, required further calibration and evaluation of this assembly protocol.

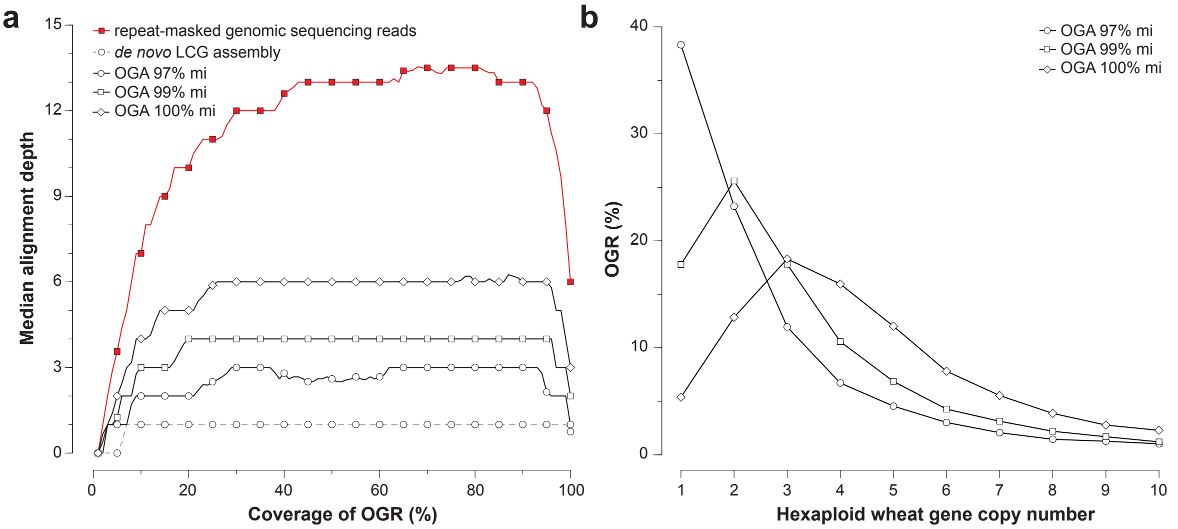


Fig. 3.3. Alignment depth of OGRs by wheat genomic shotgun reads and sub-assemblies and calculated gene copy number for different assembly parameters.
a, Median alignment depth over protein-coding regions of OGRs obtained for the repeat-masked wheat shotgun sequences, the LCG assembly and OGAs with different mi thresholds (amino terminus = 0%; carboxy terminus = 100%). **b,** Distribution of hexaploid wheat gene copy numbers predicted predicted for different OGAs.

3.1.4 Calibration of the OGA with *in silico* simulations

The mi parameter is the major factor for the assembly quality, thus impact further quantitative gene family analysis. To optimize the OGA and to calibrate this important parameter, I performed two complementary *in silico* simulations, which were based on established reference genome sequences and gene annotations of related grass genomes. For both approaches, which are discussed below, different mi thresholds were tested under consideration of the complex genome structure, the high level of repetitive sequences as well as the redundant sequences origin in three highly similar homoeologous genomes.

Simulation of a whole genome shotgun experiment
The first evaluation experiment simulated a whole genome shotgun experiment based on the maize (*Zea mays*) genome. Similar to bread wheat, maize belongs to the grass family *Poaceae* (Section 1.2), has a large genome size (approximately 2.3 Gb) and contains an extensive degree of structural complexity (several genome duplications including an ancient tetraploid state) as well as a high fraction of repeat sequences (approximately 85%) *(41)*. In 2009, the draft maize genome sequence and gene annotation were released *(41)* providing a ”gold standard” reference for evaluation of the OGA when applied to WGS data obtained for a highly complex genome with similar sequence characteristics.

This approach comprised two steps (Fig. 3.4). First, the maize gene set was catalogued and the maize gene family sizes determined, which then were paired with the previously defined orthologous group representatives. These values constituted references to compare the observed gene copy numbers in the OGA with the expected counts as annotated in the released draft genome sequence and, finally, to calibrate of the mi assembly parameter. The repeat-masked maize genome sequence (2.1 Gb) and the corresponding gene annotation incorporating 39,656 protein-coding genes *(41)* (ZMb73 version 5b.60) were retrieved and the maize proteins clustered with the protein sequences, which were also used for the definition of the orthologous grass gene set by using OrthoMCL *(223)*. This strategy identified a total of 23,086 orthologous clusters containing 28,751 (73%) maize genes. Maize gene copies with highly-similar coding sequences and sequence identity higher than the applied mi assembly parameters can not be discriminated in the OGA and would result in underestimation of the gene copy number predictions. Thus, a stringent nucleotide sequence similarity filtering of the maize gene set was additionally undertaken. Accordingly to the tested mi parameters, the maize cDNAs were clustered by using Cd-Hit *(234)* with 97%, 99% and 100% sequence similarity thresholds (parameters: -n 8 -c 0.97/0.99/1.00) and only the longest sequence of each cluster was retained as representative. Afterwards, the orthologous groups suitable for evaluation of the gene copy number were determined as defined by unambiguous clustering of OGRs (i.e. exactly one OGR per group) with one or more maize protein(s). This resulted in the selection of 15,134 (97% redundancy clustering) to 15,148 (100% redundancy clustering) groups, for which the reference gene family sizes in the maize genome was determined by counting the number of grouped maize proteins. These values were assigned to the respective OGRs.

Secondly, the OGA was applied on sequencing reads, which were computationally generated from the maize draft genome sequence using MetaSim *(235)* (version 0.9.5). The artificial reads simulated a whole genome sequencing experiment with five-fold genome coverage, sequencing error rate of 0.5% and the same empirical read length distribution as found for the wheat genomic shotgun reads. Then, repetitive sequences were filtered, the remaining reads allocated to the OGRs and maize consensus sub-assemblies generated. Thereby, I adapted the alignment parameters for the evolutionary distances between maize and the OGRs ($\geq$70% sequence identity against barley, $\geq$67% against *Brachypodium*, $\geq$64% against rice and $\geq$68% against sorghum). Three runs of the OGA were performed, each with a different assembly stringency (97% mi, 99% mi and 100% mi). Additionally, the maize gene copy numbers were computed for each OGR and each assembly iteration.

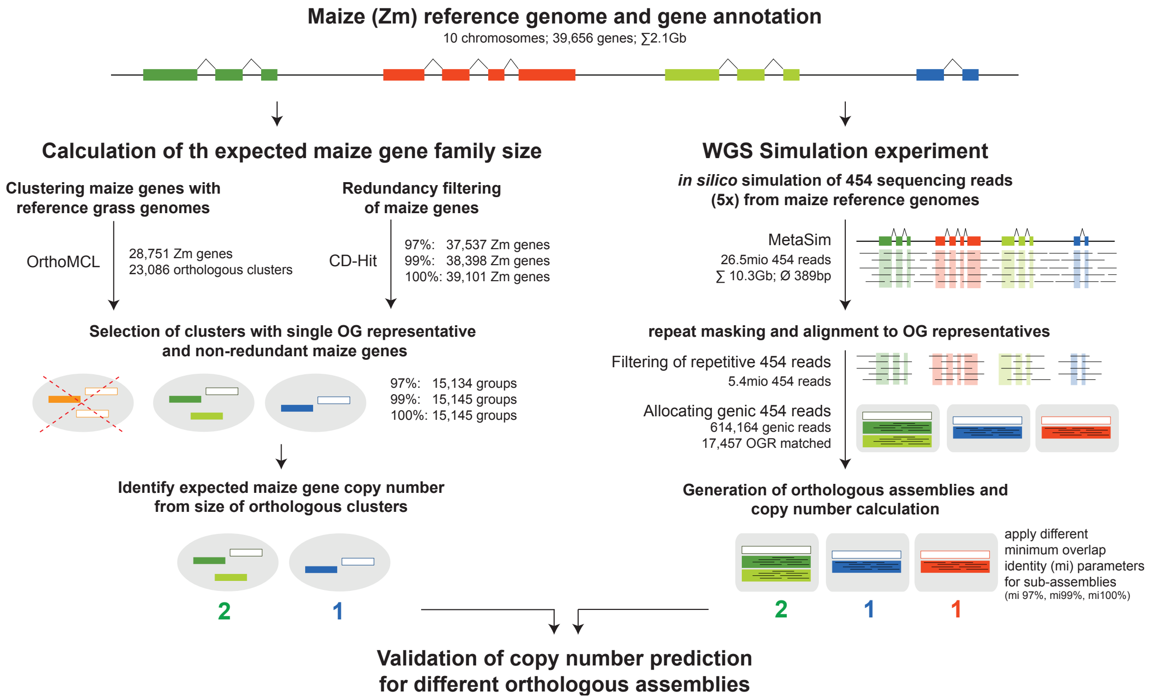

Maize (Zm) reference genome and gene annotation
10 chromosomes; 39,656 genes; ∑2.1Gb
Calculation of th expected maize gene family size
Clustering maize genes with reference grass genomes
OrthoMCL
28,751 Zm genes
23,086 orthologous clusters
Redundancy filtering of maize genes
CD-Hit
97%: 37,537 Zm genes
99%: 38,398 Zm genes
100%: 39,101 Zm genes
Selection of clusters with single OG representative and non-redundant maize genes
97%: 15,134 groups
99%: 15,145 groups
100%: 15,145 groups
Identify expected maize gene copy number from size of orthologous clusters
2
1
WGS Simulation experiment
in silico simulation of 454 sequencing reads (5x) from maize reference genomes
MetaSim
26.5mio 454 reads
∑ 10.3Gb; Ø 389bp
repeat masking and alignment to OG representatives
Filtering of repetitive 454 reads
5.4mio 454 reads
Allocating genic 454 reads
614,164 genic reads
17,457 OGR matched
Generation of orthologous assemblies and copy number calculation
apply different minimum overlap identity (mi) parameters for sub-assemblies
(mi 97%, mi99%, mi100%)
2
1
1
Validation of copy number prediction for different orthologous assemblies

Fig. 3.4. Workflow for evaluation of the OGA and the gene copy number predictions based on simulated whole genome shotgun sequencing data for the maize genome.
See text for a detailed description.

Simulation of a hexaploid gene set

Complementary, I performed a second evaluation experiment to emulate the OGA on a polyploid genome that consists of multiple, highly similar homoeologous gene copies (Fig. 3.5). Therefore, a hexaploid reference gene repertoire was computationally generated based on the annotation of protein-coding genes in rice *(45,236)*. These triplicated sequences were *in silico* "evolved" with a comparable degree as expected for bread wheat homoeologs *(123)* in order to monitor the effect

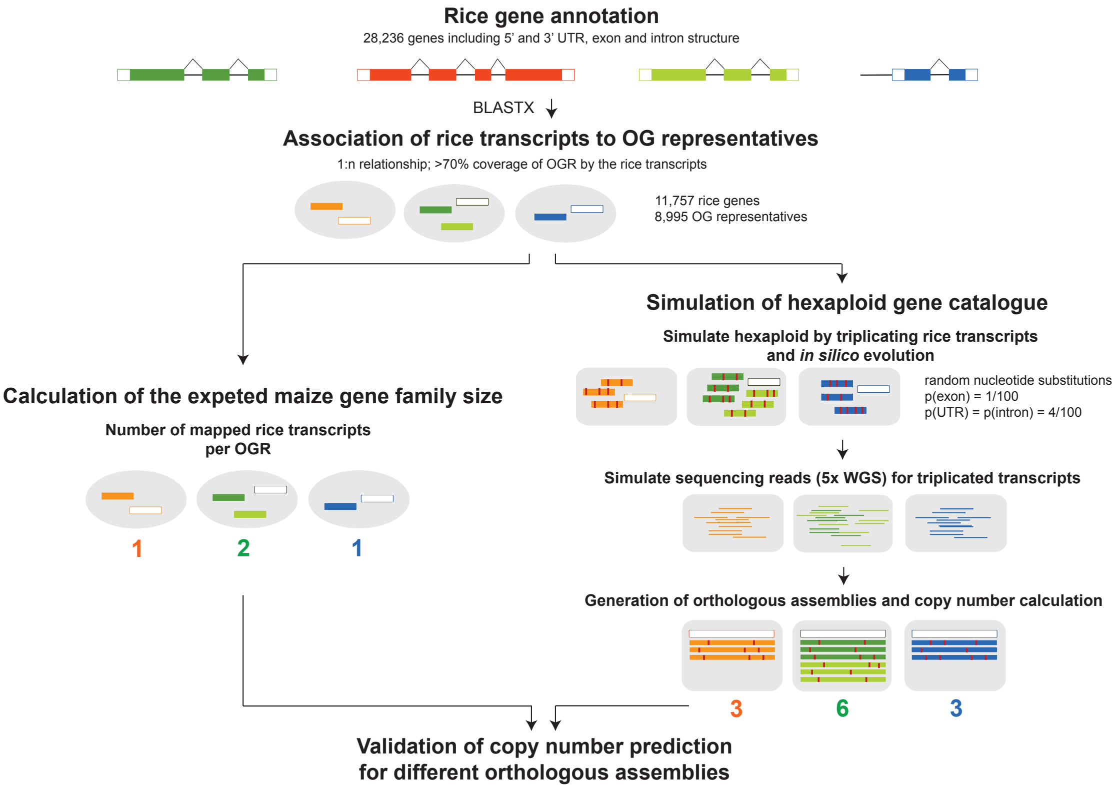

Rice gene annotation
28,236 genes including 5' and 3' UTR, exon and intron structure
BLASTX
Association of rice transcripts to OG representatives
1:n relationship; >70% coverage of OGR by the rice transcripts
11,757 rice genes
8,995 OG representatives
Calculation of the expeted maize gene family size
Number of mapped rice transcripts per OGR
1
2
1
Simulation of hexaploid gene catalogue
Simulate hexaploid by triplicating rice transcripts and *in silico* evolution
random nucleotide substitutions
p(exon) = 1/100
p(UTR) = p(intron) = 4/100
Simulate sequencing reads (5x WGS) for triplicated transcripts
Generation of orthologous assemblies and copy number calculation
3
6
3
Validation of copy number prediction for different orthologous assemblies

Fig. 3.5. Workflow for evaluation of the OGA and the gene copy number predictions based on an artificially created hexaploid gene catalogue of rice.

of high sequence similarity among homoeologs on the OGA and the copy number predictions. In detail, complete locus sequences of 28,236 rice gene models *(45,236)* (version RAP2), which were composed of coding sequences, 5' and 3'untranslated regions (UTRs) and introns, were aligned against the previously defined OGRs by using BLASTX *(227)* and adapted sequence identity thresholds (first-best blast hit with $\geq$65% sequence identity against barley, $\geq$65% against *Brachypodium* OGRs, $\geq$80% against rice OGRs and $\geq$55% against sorghum OGRs and at least 30 aa alignment length). Considering only high-confidence alignments with at least 70% coverage of an OGR, a total of 11,757 rice transcripts mapped to 8,995 OGRs. Then, the expected gene copy number in rice was determined by counting the number of aligned rice sequences per OGR.

To generate a hexaploid gene repertoire the set of aligned rice transcripts was triplicated and the divergence of homoeolog sequence copies simulated. Therefore, single nucleotides were randomly mutated with a probability of 1% (one nucleotide change per 100 bp) in protein-coding sequence regions and 4% (four nucleotide exchanges per 100 bp) in non-protein-coding sequence regions, respectively *(123)*. From these sequences artificial 454-like shotgun reads were generated with an expected five-fold genome coverage and 0.5% sequencing error rate by using MetaSim *(235)*. The obtained *in silico* reads were aligned against the corresponding OGRs [BLASTX *(227)* with *E* value $\leq 10^{-10}$ and same alignment thresholds as for the rice transcripts], sub-assembled with 97% mi, 99% mi and 100% mi and the gene copy number predicted for each

OGR and each assembly stringency.

Assessment of different assembly parameters

To evaluate the OGA approach I compared the predicted gene copy number with the gene family sizes determined by the two simulation experiments (Fig. 3.6). OGAs with 100% mi clearly exceeded the estimated one-to-one relationship for diploid maize (mean ratio of polynomial median fit 1.42) and one-to-three relationship for triplicated rice (mean ratio of polynomial median fit 7.51). When requiring perfect alignments between overlapping reads, sequencing errors predominated the OGA and prevented assembly of erroneous reads. This was consistent with the high number of singleton reads and the increased alignment depth of OGRs, which was reported previously for this setting (Table 3.2 and Fig. 3.3). On the contrary, applying minimum overlap identity of 97% underestimated substantially the gene copy numbers and resulted in a mean ratio of polynomial

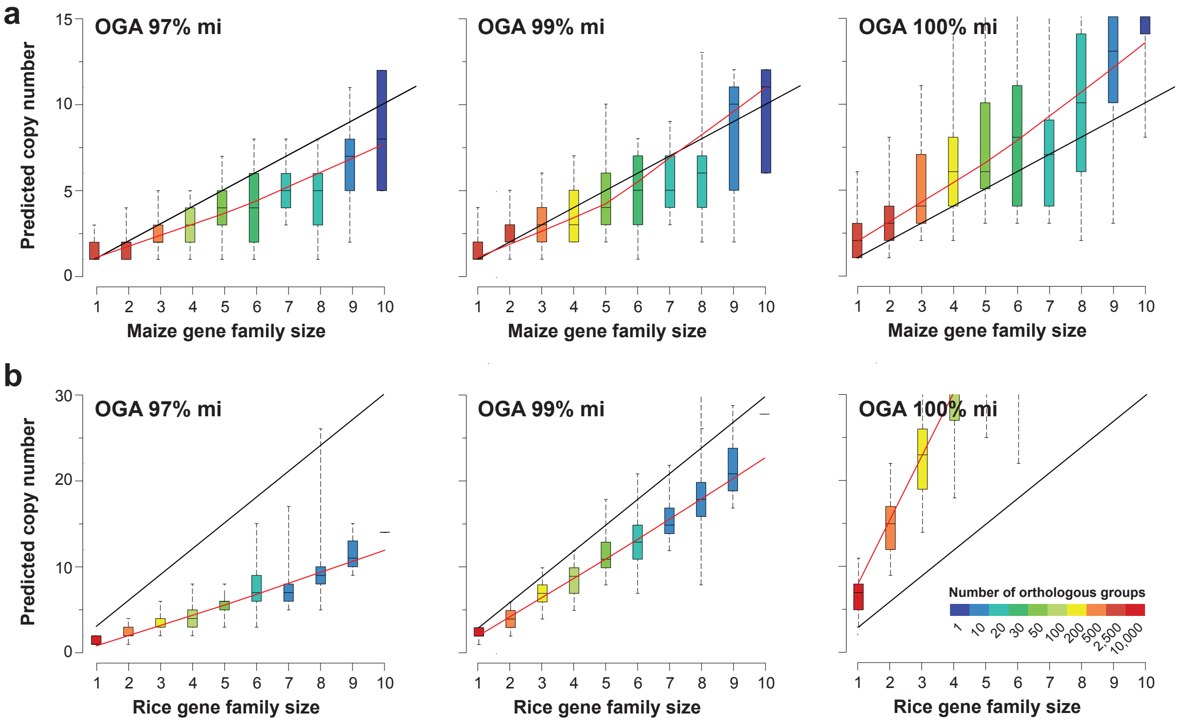


Fig. 3.6. Relationship between observed and predicted gene copy number for the simulation experiments.

The plots show the observed reference gene copy number compared against the predicted gene copy number for different orthologous group assemblies using minimum overlap identity (mi) of 97%, 99% and 100%, respectively. For each reference copy number the boxes and wiskers contain 50% and 90% of the orthologous group assembly genes, respectively. Box colors indicate the number of genes for a given copy number. The black lines represent expected gene copy numbers and the red lines show the predicted gene copy determined from the orthologous group assembly, derived by a polynomial regression fit. Only groups up to ten members are shown. **a,** Maize gene family sizes predicted from orthologous assembly of simulated genomic sequencing reads. **b,** Gene copy number predicted from orthologous assembly of simulated genomic sequencing reads derived from triplicated rice genes.

median fit 0.79 for diploid maize and 1.09 for triplicated rice.

In agreement with the observed alignment depth along the protein-coding-sequences of OGRs and the genome-wide gene family size distribution (Fig. 3.3), 99% mi outperformed the other tested assembly parameters. In both simulations, the predicted gene counts reached most closely the real gene family size distributions. Almost an one-to-one relationship between expected and predicted copy number was observed for the maize simulation (mean ratio of polynomial median fit 0.97) and the best approximation of the one-to-three relationship for ”hexaploid” rice (mean ratio of polynomial median fit 2.02). However, absence of sequence polymorphisms discriminating homoeologous gene copies caused local collapse of highly similar reads during the assembly of the triplicated rice gene set. Therefore, the predicted gene copy numbers were likely to be underestimated. Nevertheless, using this parametrization predicted the correct gene copy number within an interval of plus and minus one copy for three quarters of the OGRs.

Generally, the results highlighted that already small scale changes in the mi parameter largely affected the OGA and the predicted gene copy numbers. Consequently, a key requirement for the entire analysis was the selection of the best possible settings, which influenced significantly the interpretation of the results. Both simulation experiments showed that using 99% mi would result in a sequence assembly with the most accurate gene copy number predictions for hexaploid bread wheat. This set-up allowed compensating for sequencing errors by simultaneously maintaining distinct copies that share high sequence similarity in coding regions. Still, highly similar or identical gene copies, especially from multi-copy gene families, may have been collapsed into single assemblies and implied reduced accuracy in estimating the copy number. Therefore, the statistical analysis was restricted to gene families with maximum ten members in the diploid reference genomes.

3.2 Genome dynamics in diploid and hexaploid wheat

The previous section has demonstrated that the orthologous group assembly was suitable for the comprehensive assembly of the gene space of hexaploid bread wheat. Therefore, the presented strategy constitutes a cost-efficient method to analyse other complex plant genomes on basis of low- and medium-coverage WGS data set produced with Roche 454 pyrosequencing technology. The following section will be focussed on the investigation of the gene catalogue of bread wheat, in particular, considering the genome dynamics that happened since hybridization of the wheat lineages and subsequent cultivation. In addition to the bread wheat data set, the OGA was applied on WGS sequence reads obtained for *Ae. tauschii* (Section 2.1), the diploid progenitor of the bread wheat D genome *(67)*. The predicted gene copy numbers in *Ae. tauschii* were used as bridge to the ancestral genome state of the diploid progenitors, which enabled to elucidate gene gain, loss and duplication in the *Triticum* and *Aegilops* lineages and to monitor genome dynamics following polyploidization.

3.2.1 Orthologous group assemblies for *Ae. tauschii* and bread wheat

A similar a degree of coverage of the orthologous gene representatives was observed for simulated maize and rice shotgun sequences, experimental shotgun sequencing reads of hexaploid wheat and diploid *Ae. tauschii* and the generated sub-assemblies (Fig. 3.7a). This indicated a high comparability between the different data sets. The simulated 454 shotgun sequences of triplicated rice and the experimental sequencing reads obtained for bread wheat followed the same alignment depth distribution suggesting that the experimental data was a suitable representation of a polyploid gene catalogue sequenced with five-fold coverage (Fig. 3.7b). Contrary, simulated maize reads and experimental reads for diploid *Ae. tauschii* covered the OGRs with a median depth of five and three, consistent a diploid gene repertoire and an expected sequencing coverage of five-fold and three-fold, respectively. Both, comparable levels of coverage and alignment depth over the protein-coding regions further corroborated the previous results suggesting that the orthologous gene representatives provided a suitable proxy for comparative analysis of the diploid and hexaploid wheat genomes.

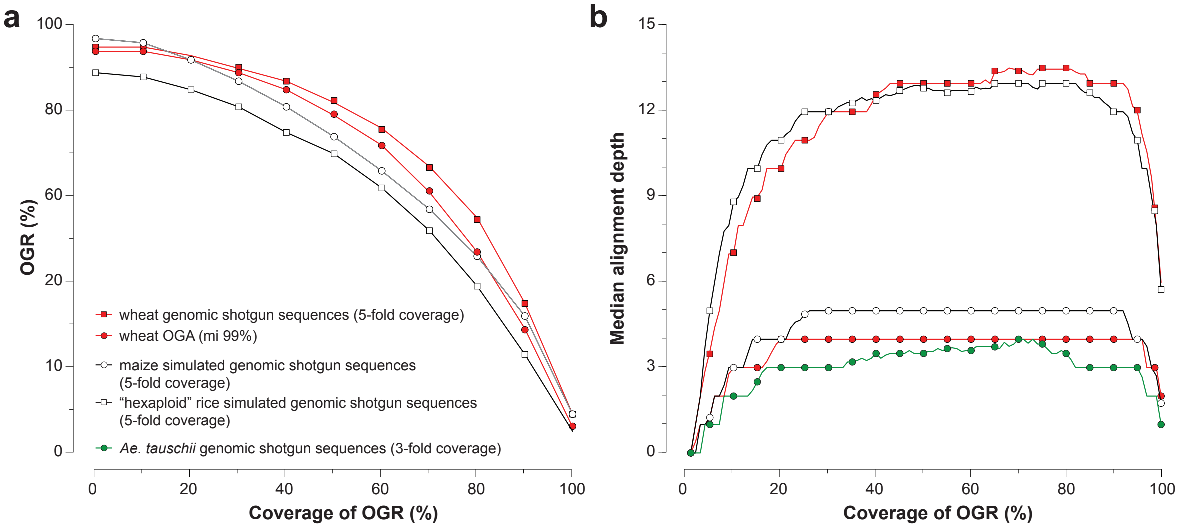


Fig. 3.7. Coverage of orthologous group representatives by raw sequencing reads and sub-assemblies.
a, Cumulative coverage of OGRs by repeat-masked 454 sequencing reads of bread wheat and *Ae. tauschii* and simulated sequences from maize and hexaploid rice. **b,** Median alignment depth over protein-coding regions of OGRs (amino terminus = 0%; carboxyl terminus = 100%).

3.2.2 Distribution of gene family sizes in wheat genomes

To investigate the impact of polyploidization on the gene content of hexaploid wheat, I determined and compared the gene copy number distributions between diploid *Ae. tauschii* and hexaploid bread wheat. For both genomes the gene copy numbers were predicted based on OGAs with 99% minimum overlap identity, which has been shown to measure gene family sizes most accurately

(Fig. 3.6). Then, I paired the observed gene family sizes in the used reference plant genome species *Brachypodium*, rice and sorghum with the predicted gene copy numbers for *Ae. tauschii* and bread wheat, respectively. Considering only OGRs with at least 70% coverage by sub-assemblies, the predicted copy number distributions were opposed to the orthologous gene family size determined for the diploid reference grass genomes.

Despite a tendency to underestimate the gene copy number for larger gene families, generally, high agreement and an almost perfect one-to-one relationship was observed between the orthologous gene family sizes in the diploid reference grasses *Brachypodium*, rice and sorghum and the predicted gene copy number in *Ae. tauschii* (Fig. 3.8a). Interestingly, high retention of homoeologous single-copy genes in hexaploid wheat was found to a similar extend as seen in *Ae. tauschii* (Fig. 3.8b). This was consistent with studies in cotton *(237)* and southern blot analyses of single-copy genes in bread wheat *(238)* suggesting only slow elimination of duplicated gene copies in small gene families. Although strong conservation of the gene family sizes was found, the results also indicated substantial variation in the gene repertoire of Triticeae genomes. In both, diploid and hexaploid wheat, numerous gene families were identified with more members than expected (genes with copy numbers above 95% confidence interval of a reference gene family size) as well as with less members than expected (genes with copy number below the 5% confidence interval of a reference gene family size). The functional biases for the expanded gene families will be investigated in more detail in Section 3.2.5. However, a general trend of

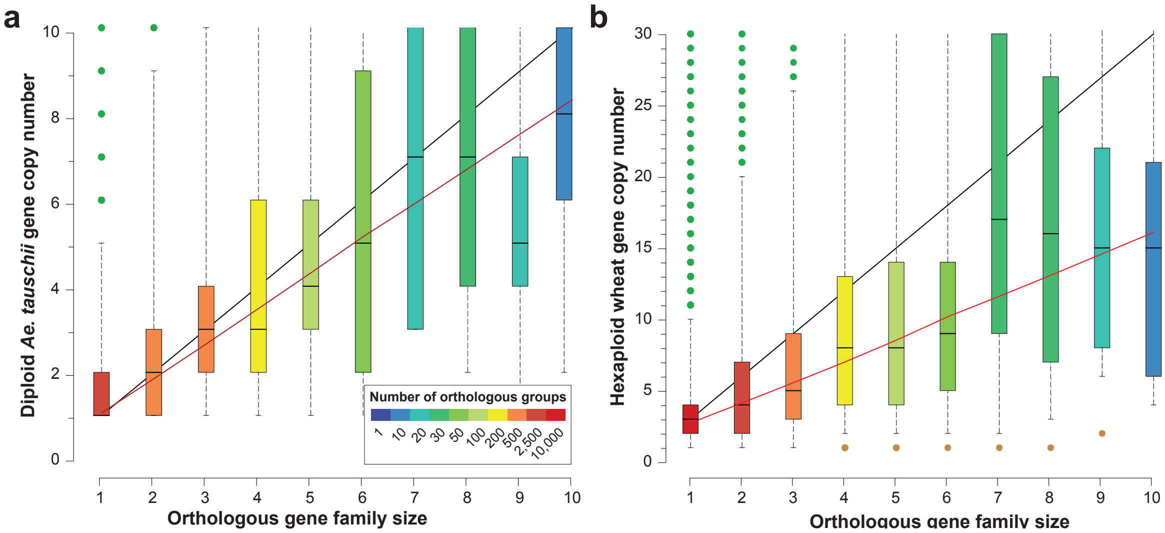


Fig. 3.8. Gene family sizes in orthologous assemblies *Ae. tauschii* and hexaploid wheat. Gene family sizes were determined by orthologous assembly of **a,** *Ae. tauschii* and **b,** hexaploid bread wheat. The boxes and whiskers contain 50% and 90% of the orthologous group assembly genes, respectively and box colors indicate the number of genes in diploid gene families of different sizes. The black lines represent expected gene family sizes, and the red lines show the gene family sizes determined from the orthologous group assembly, derived by polynomial regression fit. Only gene families with up to ten members are shown. Green dots indicate expanded gene families and brown dots contracted gene families, respectively.

gene family size reduction was apparent in bread wheat compared to the orthologous reference genomes. Thereby, the reduction was more pronounced for larger gene families, while homoeologous duplicates of single-copy genes were more likely retained. This suggested substantial loss of duplicated genes in the hexaploid genome in line with studies in bread wheat and synthetic polyploids of *Brassica* lines, which have shown that polyploids generate extensive genetic diversity by loss of DNA sequences already at an early stage of alloploydization *(47,239)*.

3.2.3 Estimation of gene number in diploid and hexaploid wheat genomes

Recent analysis of the related diploid genomes of *T. urartu* (A-genome progenitor) *(44,240)* and *Ae. tauschii* (D-genome progenitor) *(43)* (Section 1.2.4) as well as studies in bread wheat *(63,241)* have estimated each homoeologous genome to contain between 28,000 to 38,000 genes. However, the gene content of bread wheat has only been extrapolated from sequences of single chromosomes so far *(63,241)*. Based on the copy number predictions and the comparison to the orthologous gene family sizes, this work allowed to determine the gene number of the diploid *Ae. tauschii* genome ($2\mathrm{n}$) and, for the first time on a genome-wide level, the hexaploid bread wheat genome ($6\mathrm{n}$). Therefore, the observed gene retention rates ($r_{\mathrm{predicted}}$) were defined from the slopes of the polynomial regression fit of the gene family distributions of *Ae. tauschii* [$r_{predicted}(2\mathrm{n}) = 0.91$] and hexaploid bread wheat [$r_{\mathrm{predicted}}(6\mathrm{n}) = 1.83$] (Fig. 3.8). These rates were additionally corrected for technical limitations in the estimation of the gene copy number, which caused underestimation of gene counts by partial collapse of highly similar sequences in the assemblies (Fig. 3.9). The corresponding correction factors (δ) were inferred from the deviations of the predicted to the expected copy numbers in the simulation experiments of a diploid gene set using maize [deviation to an expected one-to-one relationship; $\delta(2\mathrm{n}) = 0.97/1$] and for a hexaploid gene set from triplicated rice [deviation to an expected three-to-one relationship; $\delta(6\mathrm{n}) = 2.21/3 = 0.74$]. The corrected gene retention rate r were computed as:

$$r_{\mathrm{corrected}} = \frac{r_{\mathrm{predicted}}}{\delta} \tag{3.2}$$

This resulted in a corrected gene retention rate of 0.94 : 1 for *Ae. tauschii* and 2.48 : 1 for bread wheat, respectively.

A total of 18,508 and 58,758 distinct high-confidence copies (G_{HC}) were identified for *Ae. tauschii* and hexaploid wheat, respectively. These covered the protein-coding sequence of 7,116 (*Ae. tauschii*) and 12,481 (bread wheat) OGRs with at least 70%. For the remaining 12,885 and 7,570 low-confidence OGRs (G_{HC}) with medium- or low-coverage the gene copy number were extrapolated from the corrected gene retention rate ($r_{\mathrm{corrected}}$) and the average orthologous gene family size ($s = 1.46$) observed in the orthologous reference genomes *Brachypodium*, rice and sorghum, respectively. As determined in Section 3.1.1, considering 92% of wheat genes to be detectable by using the defined orthologous gene framework ($d = 0.92$) allowed to estimate the

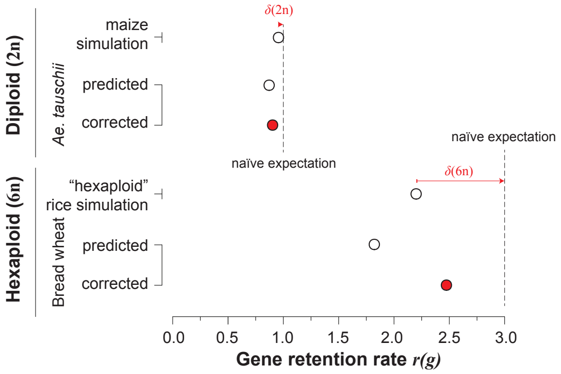


Fig. 3.9. Gene retention rates for diploid *Ae. tauschii* and hexaploid wheat.
Predicted gene retention rates of *Ae. tauschii* and bread wheat were computed and corrected for technical limitations, which were deduced from the deviations (red arrows) to the perfect prediction of gene copy numbers for diploid and hexaploid genomes (dashed lines).

total gene repertoire G for both genomes:

$$G = \frac{G_{\mathrm{HC}} + G_{\mathrm{LC}}\ r\ s}{d} \tag{3.3}$$

This calculation resulted in an estimate of 39,000 genes *Ae. tauschii* and 94,000 genes for hexaploid wheat, which was reasonable consistent with independent estimates of other studies *(43,44,63,240,241)*.

3.2.4 Genome change in polyploid wheat

To further investigate the genome change in the polyploid wheat, I directly compared the copy number distributions of *Ae. tauschii*, a proxy for the diploid wheat progenitor genomes, and bread wheat. As shown by the amalgamation of the diploid and hexaploid wheat gene copy numbers in Fig. 3.10, for all orthologous gene family sizes on average less copies were predicted in the hexaploid genome as compared to the diploid genome. The lower number of detected orthologous gene copies suggested substantial gene loss in bread wheat, which is indicated by the grey zone between the regression fits. Based on the previously calculated gene retention rates for *Ae. tauschii* and hexaploid wheat, the hexaploid-to-diploid gene family size ratio was estimated to be $2.48/0.94 : 1/1 = 2.64 : 1$. Therefore, the comparison of the observed hexaploid-to-diploid gene family size ratio with a naïvely expected ratio of $3 : 1$ allowed to estimate the loss of approximately 12,000 genes (12%) in hexaploid wheat compared to the ancestral diploid progenitor genomes. This estimate was largely consistent with earlier studies of gene loss in newly synthesized wheat polyploids *(242)* and the erosion of genetic diversity during domestication *(48)*. Moreover, the predictions corroborated recent estimates of Dvorak *et al. (243)*, who detected 26 out of 155 investigated loci (17%) to be deleted during the evolution of polyploid wheat by hybridization mapping of expressed sequence tags with bread wheat deletion stocks. This study substantially increased comprehensiveness and resolution compared to previous analysis and, therefore, allowed further genome-scale monitoring with robust statistical testing and analysis for functional implications of polyploidy.

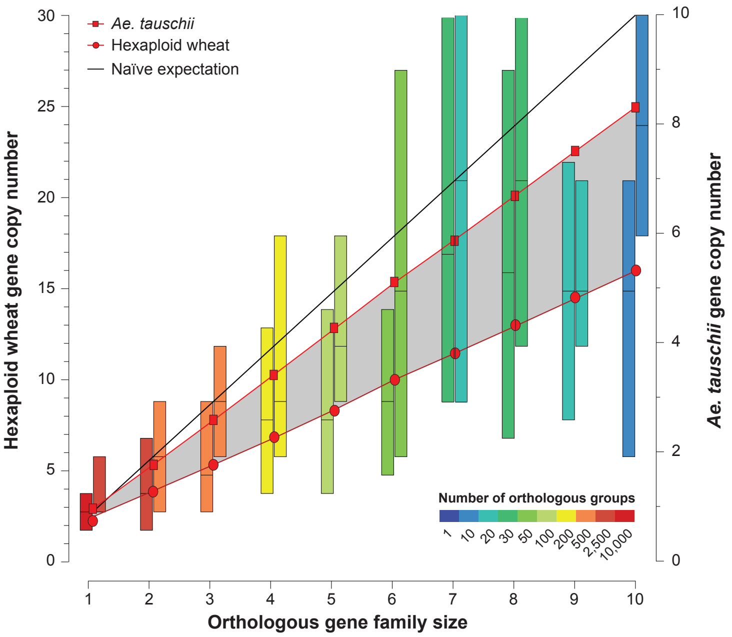


Fig. 3.10. Amalgamation of diploid and hexaploid wheat gene copy numbers.
Observed gene copy number of bread wheat and *Ae. tauschii* for respective orthologous gene family sizes. The boxes contain 50% (lower and upper quantiles) of the orthologous group assembly genes. The black line indicates the expected gene family sizes (one-to-one for *Ae. tauschii* and three-to-one for hexaploid wheat, respectively). Red lines show the polynomial regression fit of observed copy numbers. The grey zone between the regression lines estimates the extent of gene loss in hexaploid wheat. For each family size, the left-hand boxes represent hexaploid wheat and right-hand boxes represent *Ae. tauschii*.

3.2.5 Functional analysis of expanded gene families in *Ae. tauschii* and bread wheat

Loss, retention and amplification of gene copies may influence the proteome in various ways *(244)*. On the one hand, recent studies have shown deleterious effects of gene duplications and identified genes that were convergently restored to singleton status following polyploidization *(245,246)*. On the other hand, duplicated genes might retain in the genome and preserve their original gene functions *(247)* or provide a redundant gene pool allowing the development of new functionalities with strong advantageous effects on a species' fitness *(248)*. Thus, genome duplications and polyploidzation may constitute beneficially to, for example, the adaption of to changing environments *(249)*.

In this study, various gene families were identified with expanded copy numbers in diploid *Ae. tauschii* as well as in hexaploid wheat (Fig. 3.8, green dots). To further test for functional implications of gene family expansion, the gene ontology (GO) categories were inferred for these genes from the cognate *Brachypodium*, rice, sorghum and barley OGRs as annotated in "The Similarity Matrix of Protein" database *(250)*. Significant over-represented ontologies were determined by functional enrichment analyses, which were independently performed for expanded *Ae. tauschii* and bread wheat gene families[(4)]. Subsequently, the identified over-represented GO terms were opposed to each other (P <0.05).

A large fraction of the significantly over-represented functional categories were shared between *Ae. tauschii* and hexaploid wheat (Fig. 3.11). These included, for example, proteins related to "manganese ion binding" *(251)*, "flavin-containing monooxygenase activity", "nicotinamide adenine dinucleotide dehydrogenase activity" *(252)* or "oxidoreductase activity" *(253)*, which suggested expansion of gene families for basal cellular reactions and developmental processes in Triticeae genomes. Genes encoding for components involved in photosynthesis [e.g. "chlorophyll binding" or "electron carrier activity" *(254)*] as well as genes function in immune and defence responses as well as resistance against pathogen invasion [e.g. "MHC class I protein binding", "chitin binding", "cysteine-type endopeptidase activity" *(255,256)*] were also more abundant in *Ae. tauschii* and hexaploid wheat compared to other closely related grasses.

On the contrary, some molecular functions were found to be exclusively expanded for *Ae. tauschii*. For example, gene families encoding for hydrogen ion transmembrane transporters and different subunits of ATPases ("proton-transporting ATPase activity") may provide proton gradients to support Na^+ exclusion in *Ae. tauschii (257)* and the accumulation of minerals in other *Aegilops* species *(258)*. Vice versa, proteins involved in the responses to biotic and abiotic stresses [e.g. "pattern binding" *(259)*, "methyl jasmonate esterase activity" *(260)* or "methyl salicylate esterase activity" *(261,262)*] were found to be expanded in bread wheat only. Additionally, an increased gene copy number in bread wheat was also found processes related to seed and storage compounds like, for example, "protein tyrosine kinase activity", which is involved in the mobilization of seed proteins and lipid reserves *(263)*, or "glutathione transferase activity", which is important for grain filling and embryo development *(264)*.

In summary, these observations showed that gene families related to essential molecular processes and functions are commonly expanded for the Triticeae *(265,266)* indicating that at least part of the genetic characteristics of bread wheat were already defined in the diploid progenitor genome(s), inherited to hexaploid wheat and maintained during polyploid evolution. However, many gene families of agricultural importance were exclusively expanded in bread wheat. Assuming that those observed gene family expansions not origin in the diploid genomes of the other progenitors for the wheat A- and B-genome, which could not be excluded with the current data set, this finding suggested that selection during domestication might have contributed to the expansion of agriculturally important gene families in the bread wheat genome.

[(4)] I gratefully acknowledge and thank co-author Manuel Spannagl, who implemented and performed the functional enrichment tests based on the selected gene families with expanded sizes in *Ae. tauschii* and bread wheat, respectively.

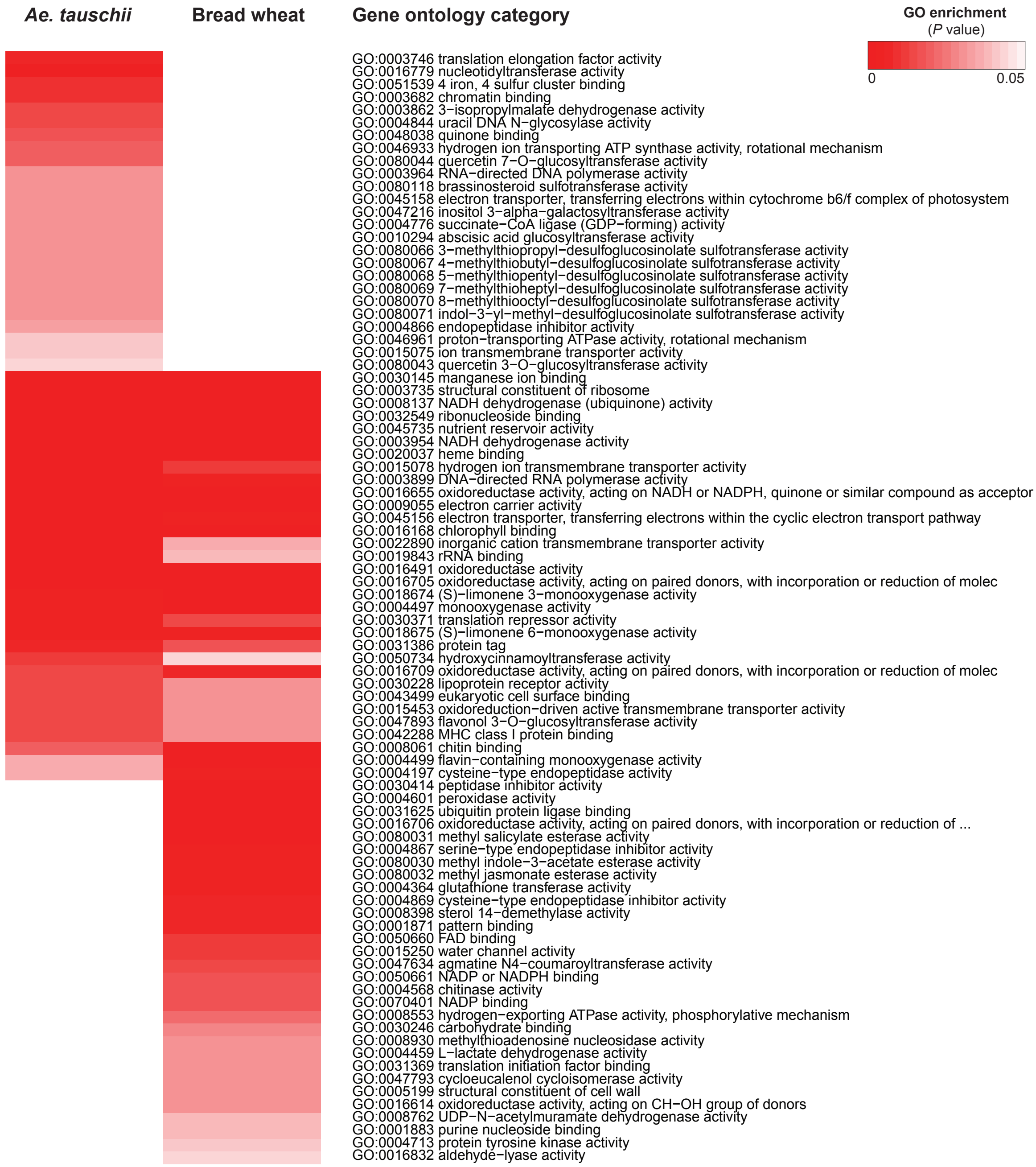


Fig. 3.11. Significant over-represented gene ontology categories of expanded gene families in *Ae. tauschii* and hexaploid wheat.
Orthologous group representatives were identified that had significant elevated copy number in *Ae. tauschii* or in bread wheat and were subjected to functional analysis using GO enrichment test. All significant over-represented molecular functions with *P* values <0.05 are shown.

3.3 Signatures of pseudogenes in the wheat genome

In 1977 Jacq and co-workers identified a truncated and not expressed copy of the 5S ribosomal RNA gene in *Xenopus laevis (267)*. They characterised this genomic sequence to be most probable ”a relict of evolution” and, hence, termed it ”pseudogene”. Ever since then pseudogenes have been detected in almost all analysed genomes within the three kingdoms of life *(268–270)*. Pseudogenes have been defined as genomic sequences derived from a functional RNA or protein-coding gene, which lost their potential to encode for functional products *(271,272)*. They exhibit substantial sequence similarity to a functional gene, but also degenerative sequences features are present, such as truncations of the full-length gene or deleterious mutations resulting in premature stops and frame-shifts *(272–274)*. Based on the underlying causative mechanism pseudogenes can be classified into two major groups: processed pseudogenes, which are derived from duplication of genomic DNA by whole genome, tandem or segmental duplications and non-processed pseudogenes, which origin from retro-transposition of a RNA intermediate back into the genome *(270)*. Several classes of plant DNA transposons *(224,275)* and retroelements *(275)* as well as the double-strand break repair mechanism *(199,225)* cause and amplify gene fragments and have been discussed to disrupt genes and generate pseudogenes.

The role and function of pseudogenes is not entirely understood. While pseudogenes have been shown to evolve neutrally *(276)* and are by definition non-functional at the protein level, recent studies have demonstrated that some pseudogenes are expressed *(277)* and potentially exert regulatory functions *(270,278)*. However, independently from their functional relevance, pseudogenes may provide a reservoir of genetic diversity supporting the evolution of new genes and contribute to the formation of gene families *(274,279)*.

Manual inspection of alignments between wheat sub-assemblies against the cognate orthologous group representatives revealed frequent occurrence of local ”stacks” of gene fragments, which were aligned to the same protein-coding region of an OGR (Fig. 3.12). These stacks comprised several distinct sub-assemblies, which were sufficiently divergent not to assemble. While intact gene assemblies would cover the almost entire protein sequence of the OGRs, increased alignment depth covering only a local segment of a protein-coding gene might origin from amplification of genomic sequences by pseudogene-causing mechanisms like, for example, retro-transposition of RNA intermediates back into the genome *(199)*. The following sections will specifically investigate the formation of pseudogenes in the bread wheat genome and discuss and characterize the sequences forming local stacks.

3.3.1 Identification of pseudogene candidates

Local stacks were systematically identified based on the relative mapping depth along the protein-coding sequence of each OGR. This measure was defined and calculated by normalizing the hit count profile (i.e. the number of aligned sub-assemblies per amino acid) to the previously pre-

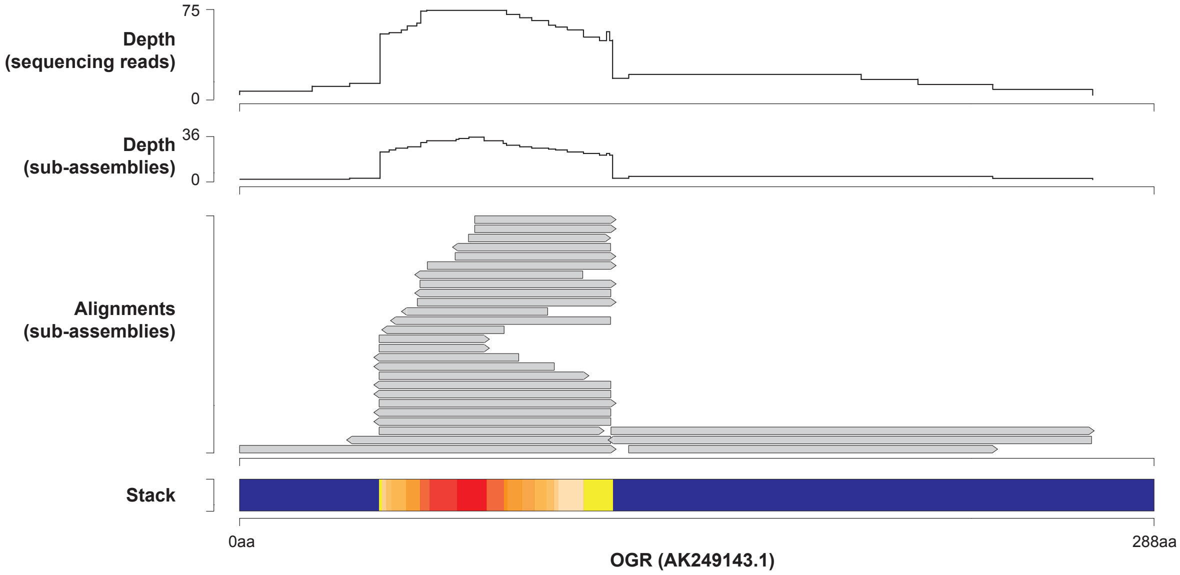


Fig. 3.12. Example of an OGR with associated wheat sub-assemblies and a "stack" region. Visualization of the alignment depth of repeat-masked genomic shotgun sequencing reads (top track) and wheat sub-assembly sequences (second track) along the protein sequence of an OGR. Alignments for5 wheat sub-assemblies are shown. The heat map depicts the protein-region of stacked gene fragments.

dicted wheat gene copy number. Stacks were defined as protein-coding regions of OGRs, which showed at least five-fold increase in the number of aligned sub-assemblies relative to the predicted copy number over a continuous stretch of minimum 30 amino acids (Fig. 3.12). These were further categorized into two types. Stacks of the first type overlapped with a known protein family (Pfam) domain *(280)* of the orthologous group representative, thus were termed "Pfam-stacks". Since these stacks were associated to well-conserved protein domains, they might have originated from genomic sequence reads of related genes, which were absent in the orthologous gene framework used for the OGA, but shared fractional sequence homology. The second type of stack was not overlapping with any Pfam domain and, due to their multiple fragmentary composition, this type of stacks was termed "pseudogene-stacks".

A total of 5,538 stacks were identified for 3,648 OGRs (29%), which were covered at least 70% by wheat sub-assemblies (Table 3.3). The majority of these were classified as pseudogene-stacks (72%). Furthermore, almost one third of the sub-assemblies (232,877) were detected to overlap by at least 90% of their sequence with an identified stack regions (Table 3.3). A total of 162,930 sub-assemblies (21%) were contained in pseudogene-stacks. This observed proportion was largely consistent with the classification of 27 pseudogenes out of 148 predicted gene candidates (18%) in the analysis of 13 Mb-sized BAC contig sequences of chromosome 3B *(241)*. The identified gene fragments had a mean length of 165 bp and most of the stacks covered between 5% to 15% of an OGR's length (Fig. 3.13a). Corroborating previous studies, the reduced coverage of genes by stacks indicated that these sub-assemblies might represent gene fragments originating from transposable element capturing and double strand break repair mechanisms *(199,224,225,275)*. On average, stacks were present with nine-fold greater depth than the predicted gene copy number. A strong trend for pseudogene stacks to be preferentially located at the terminal regions of the orthologous group representatives was evident and contrasted with the distribution of Pfam-stacks, which were found equally located across the protein-coding sequences of OGRs (Fig. 3.13b).

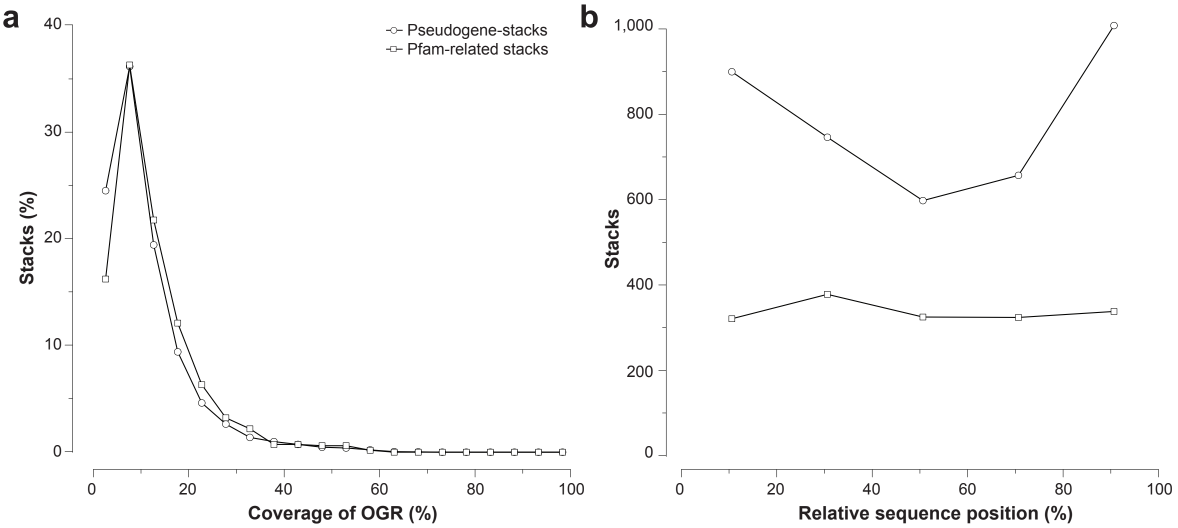


Fig. 3.13. Gene coverage and localization of identified stack regions.
a, Frequency distribution of identified stack regions relative to their coverage of protein-sequence of the OGRs. **b,** Localization of identified stacks compared to their relative sequence position in the protein-sequence of the cognate OGRs (amino terminus = 0%; carboxy terminus = 100%). The protein-coding sequence of each OGR was divided into four segments and the number of stacks located within each segment was counted.

Table 3.3. Analysis of of gene fragments and sub-assemblies forming local stacks.
Local stacks were identified based on the relative exceed in alignment depth compared to the predicted gene copy number. Only well-covered orthologous group representatives (i.e. ≥70% coverage by sub-assemblies) were considered in this analysis.

	Pfam-stacks	Pseudogene-stacks	Total[(a)]
Analysed OGRs (≥70% coverage)	-	-	12,518
Analysed sub-assemblies	-	-	761,470
Identified stacks	1,661	3,877	5,538
OGR with stacks	1,266 (10%)	2,631 (21%)	3,648 (29%)
Su-bassemblies associated to stacks	69,947 (9%)	162,930 (21%)	232,877 (31%)
Mean coverage of OGR by stacks	12.19%	10.85%	11.25%
Mean length of stacks	171bp	163bp	165bp
Mean depth of stacks[(b)]	35.64	32.51	33.45
Mean exceed of depth[(c)]	9.43	8.79	8.98

[(a)] Orthologous group representatives including Pfam-related and "pseudogene" stacks were counted once.
[(b)] Depth measured as number of aligned sub-assemblies at a sequence position of the OGR.
[(c)] Exceed of depth was calculated as the mean ratio of the alignment depth[(b)] compared to the predicted gene copy number of an OGR.

3.3.2 Signatures of selection pressure on stack sub-assemblies

By definition, pseudogenes are released from functional constraints and expected to evolve neutrally *(271)*. Those changes either promote functional divergence or lead to inactivation and silencing of the gene. To gain insight in the evolutionary fate of the detected wheat sequences forming stacks, pairwise protein alignments were computed and analysed between the wheat sub-assemblies and their cognate orthologous group representative. Sequence conservation of sub-assemblies in Pfam- and pseudogene-stacks decreased approximately ten percent compared to those sequences, which were not associated to any stack (Fig. 3.14a). This significant decrease in protein similarity indicated substantial divergence in protein sequence of wheat sub-assemblies forming stacks compared to sub-assemblies representing intact gene copies [Wilcoxon-Mann-Whitney-Test (*P* <10^{-20})].

Furthermore, I performed a sequence divergence analysis to elucidate the relationship between the coverage of OGR by individual sub-assemblies and the evolutionary constraints on those sequences. Therefore, the number of nonsynonymous substitutions per nonsynonymous site (K_a) and the number of synonymous substitutions per synonymous site (K_s) were determined between sub-assemblies and the respective orthologous group representatives. The ratio of both values (K_a/K_s) measures the strength of selection acting on the assembled wheat sequences. Values below one indicate stabilizing or purifying selection ($K_a/K_s < 1$), values greater than one indicate positive selection ($K_a/K_s > 1$) and values around one suggest neutral selection

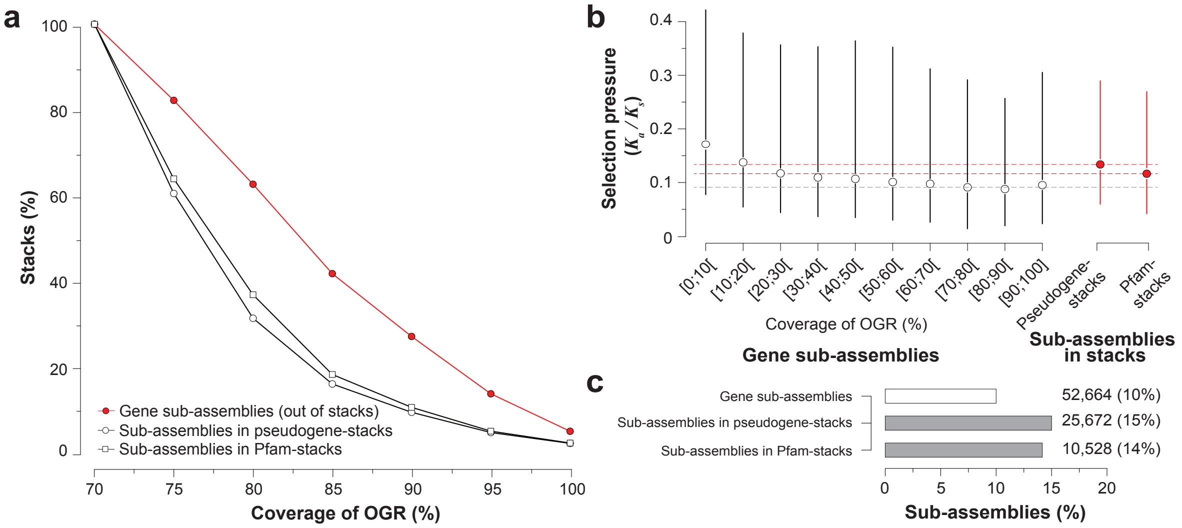


Fig. 3.14. Sequence diversity analysis of wheat sub-assemblies in stacks and non-stack regions. **a,** Cumulative frequency distribution of protein alignment similarity between the predicted protein sequence of sub-assemblies and the respective OGRs. **b,** Identified sub-assemblies with disrupted protein sequences by premature stop codons. **c,** K_a/K_s distribution for sub-assemblies that are out of stacks and assigned to pseudogene- and Pfam-stacks, respectively. For gene sub-assemblies (out of stacks), the K_a/K_s distribution is shown respective to the coverage of the OGR. Only single exon OGRs were considered.

($K_a/K_s = 1$). Generally, the OGA was limited in the assembly of sequences bridging intros and complete gene sequences could only be assembled for single exon OGRs (Fig. 3.1). Therefore, the subsequent analysis was restricted to OGRs, which protein-sequence was encoded by a single exon, thus could have been assembled in full-length. The computed protein sequence alignments between OGR and sub-assemblies were translated into the corresponding nucleotide alignments and the K_s, K_a and K_a/K_s rates determined by using the yn00-tool implemented in the PAML4 package *(281,282)*.

Significant differences in the K_a/K_s distributions were evident with respect to the coverage of the OGR by gene sub-assemblies [Wilcoxon-Mann-Whitney-Test (*P* <0.01)] (Fig. 3.14b). K_a/K_s values of sub-assemblies that cover only up to 20% of the OGR were significantly increased compared to those covering the OGRs in full-length. This indicated that sub-assemblies covering only local regions of their cognate OGR were less functionally constrained and suggested these fragments were biased towards neutral evolution. However, mean K_a/K_s values lower than 1 were observed. The deviation from this expected value for pseudogenes, which would indicate complete release from any functional restraint, can be explained in two ways. First, the wheat sub-assemblies were compared against the orthologous genes in *Brachypodium*, rice, sorghum or barley, rather then to the true parental (wheat) gene. The sequence evolution of the relatively distant reference genes resulted in increased K_s values, which lowered the K_a/K_s ratios. Secondly, this work focussed only on relatively young pseudogene candidates due to the stringent alignment thresholds used for the alignment of wheat sequences against the OGRs. Thus, the considered evolutionary time frame was too short and only small changes in the K_a/K_s were observable.

Accumulation of nucleotide substitutions causing premature stop codons are the most radical changes, which truncate and disrupt the encoded amino acid sequences and result most probable in inactivation and non-functional proteins. Strikingly, one third more sub-assemblies with disrupted protein sequences were associated to pseudogene-stacks (15%) and Pfam-stacks (14%) compared to sub-assemblies, which were not associated to stack regions (Fig. 3.14c). In addition to increased protein divergence and reduced functional constraints, this observation supported the hypothesis that stacks are potentially formed by wheat sequences, which resulted from generation and amplification by pseudogene-forming mechanisms. However, this study could not entirely exclude that these sequences also constituted fragments of functional genes that were not completely represented in the defined orthologous gene family framework. Vice versa, some of the sub-assemblies, which were not assigned to stacks and covered only small portion of a OGR, might also be gene fragments or pseudogenes that have not sufficiently been amplified to form stacks.

3.3.3 Over-representation of pseudogenes among domain families

To investigate the extent to which certain functional classes, gene families or traits have been affected by pseudogene formation in bread wheat, a GO enrichment test was performed comparing the Pfam domain designations of OGRs with identified pseudogene-stacks versus the entire set of all OGRs[(5)]. This analysis revealed several protein family signatures to be over-represented (Table 3.4) including proteins that are known to be involved in pseudogene formation and gene transposition and encode zinc-finger motifs in mutator transposons *(224)* and plant mobile domains *(283)*. Furthermore, genes encoding nucleotide-binding site leucine-rich repeat proteins, which are involved in plant resistance and defence response to environmental stimuli and biotic or biotic stresses *(284–286)*, members of the F-box protein family, which are important for protein-protein interactions *(287,288)*, and proteins with kinase domains were also significantly enriched for stacks. This was in large agreement with the rapid evolution and high diversification reported for these gene families *(42,137,285)* and studies in *Arabidopsis (289,290)* and rice *(289)*, which have demonstrated that especially defence gene families are affected by pseudogene formation and shaped by lineage-specific tandem duplication and subsequent selection *(291)*.

Table 3.4. Functional analysis of OGRs with pseudogene-stacks.
OGRs with identified pseudogene-stacks were subject to GO enrichment analysis. The table summarizes the over-represented Pfam domains up to a maximum Bonferroni corrected *P* value of 0.1.

PFAM-Id	*P* value	Corrected *P* value	Description
PF00646	4.0 x 10^{-0}	3.6 x 10^{-7}	F-box domain
PF04578	9.4 x 10^{-8}	8.4 x 10^{-5}	Protein of unknown function
PF00560	6.3 x 10^{-7}	5.7 x 10^{-4}	Leucine Rich Repeat (LRR)
PF00069	1.3 x 10^{-6}	1.2 x 10^{-3}	Protein kinase domain
PF10551	7.7 x 10^{-5}	6.9 x 10^{-2}	Mcl-1 ubiquitin ligase E3 (MULE) transposase domain
PF10536	7.9 x 10^{-5}	7.1 x 10^{-2}	Plant mobile domain

Although that some pseudogene-stacks might have originated from an incomplete representation of wheat genes in the utilized orthologous gene family framework or double strand break filling mechanisms using reverse transcribed mRNAs *(199,225)*, especially gene families involved in plant resistance and defence response generated stacks, thus were frequently affected by gene duplication and pseudogenization. This observation constituted an indicator for the formation of new gene functions via gene duplications mechanisms in wheat *(291)*. As proposed in the gene "birth-and-death" evolution model *(292)*, such mechanisms impose a high redundancy in the gene pool and might facilitate rapid modifications on protein sequence level providing a reservoir for selection in consequence of adaption to ever-changing environment *(291)*.

[(5)] I gratefully acknowledge Manuel Spannagl, who implemented and performed the functional enrichment test based on the list of candidate OGRs.

3.4 Conclusions

Whole genome shotgun strategies utilizing NGS technologies provide a rapid and cost-efficient way to obtain large collections of fragmented genomic DNA sequences. However, especially the assembly and analysis of the generated data sets for large, complex and, in particular, polyploid plant genomes has been a severe challenge. Based on low-coverage whole genome sequencing of the bread wheat genome, a novel bioinformatic assembly strategy was developed exploiting comparative-genomics to generate comprehensive genomic sequence resources. Comparison with the sequence of the diploid D-genome progenitor *Ae. tauschii* revealed high retention of homoeologous genes, but also a general trend in gene family size reduction, which was consistent with small-scale analyses *(48,293)*. The observed degree of gene loss in hexaploid wheat was considerably smaller as compared with paleopolyploid maize *(111)* and mesopolyploid *Brassica rapa (50)*. This might be caused by its relatively recent origin and the absence of intergenome recombination *(294)* (Section 1.3.1). However, pronounced gene loss in large gene families wheat corroborated rapid genomic changes as observed for allopolyploid *Tragopogon miscellus (295)*. This work also identified several classes of gene families with increased sizes in the *Triticum* and *Aegilops* lineages, which were linked to important agricultural and industrial characteristics of bread wheat including defence, nutritional content, energy metabolism and growth. High abundance of gene fragments, often forming ”stacks”, highlighted the plasticity of the bread wheat genome and indicated that gene duplications may have contributed to the formation of new gene functions and the rapid evolution of gene families related to environmental responses *(292)*.

Major efforts are underway to improve wheat productivity by increasing genetic diversity in breeding materials and through genetic analysis of traits *(21)*. All developed genomic resources were made public available and deposited in the European Nucleotide Archive (ENA) with project accession PRJEB568. Although the produced assemblies are fragmentary, they will constitute a framework for identification of genes, supporting further genome sequencing and facilitating genome-wide analyses. However, alternative strategies are needed to establish a (draft) reference genome sequence and structural gene annotation for the bread wheat genome. In particular, distinguishing homoeologous chromosomes prior to sequencing by using chromosome flow-sorting technology *(205–207)* constitute a powerful technologies allowing to assemble homoeologous chromosomes individually *(198)*. This complementary strategy will be further discussed in the following chapter of this thesis.

Chapter 4

A chromosomal survey of the bread wheat genome
– Gene annotation and genome analysis –

The previous chapter discussed opportunities and limitations of whole genome shotgun sequencing for the analysis of the bread wheat genome. While algorithms that allow quantification of the gene repertoire of hexaploid wheat were developed and applied to measure globally genome change following polyploidization, sequencing the whole genome at once could not be used to generate a suitable draft reference genome sequence. The majority of homoeologous sequence copies were collapsed in the genome-wide *de novo* assembly of WGS reads, which restricts genome-specific identification of full-length protein-coding sequences and structural annotation. Moreover, assignment of the genome-of-origin has been constrained only to a subset of assembled sequences, thus limiting further investigation of homoeologous relationships and phylogenetic analysis. To overcome these challenges the IWGSC employed a ”divide-and-conquer” approach and utilized flow-cytometry technology to isolate, purify and sequence DNA of individual wheat chromosome arms *(209,296)*. This strategy allowed generating a ”chromosomal survey sequence” assembly based on Illumina short reads *(297)* (Section 2.2), which constituted a valuable draft genome sequence and permitted to distinguish homoeologs and to structurally annotate the bread wheat genome.

In the following sections, the annotation and characterization of the bread wheat genome will be described. Therefore, I implemented an extrinsic gene prediction pipeline to identify protein-coding gene loci, alternative transcript usage and novel (non-protein-coding) transcriptional active regions (nTARs). For the first time, this annotation allowed large-scale comparative analysis between the wheat A, B and D genomes and investigation of syntenic conservation and gene family composition on a chromosome (arm) level. Here, structural attributes of homoeologous genes and transcripts will be described, patterns of alternative splicing elucidated and aspects of post-transcriptional gene expression regulation discussed.

In its first version the gene annotation pipeline was implemented for the annotation of the barley draft genome sequence assembly *(191)*. An adapted version, which will be described in this chapter, was applied on the CSS assembly of bread wheat.

All methods and results shown in this chapter are part of following publications:

- **A physical, genetic and functional sequence assembly of the barley genome**
  The International Barley Genome Sequencing Consortium (IBSC)
  Nature. 491(7426):711-717, 2012.

- **A chromosome-based draft sequence of the hexaploid wheat genome**
  The International Wheat Genome Sequencing Consortium (IWGSC)
  Science. 345(6194):1251788, 2014

4.1 Exon detection and consensus gene modelling

Computational approaches for the discovery and annotation of gene structures on basis of a reference genome sequence are mainly divided into two categories: *ab initio* (or intrinsic) and homology-based (or extrinsic) methods *(298)*. *Ab initio* gene finders apply statistical models to detect genes based on characteristic genomic sequence features. In contrast, homology-based approaches utilize alignments of external evidences, for example cDNA or protein sequences from closely related species, to annotate genes and corresponding structures. Hybrid approaches that combine advantageous of intrinsic (no additional data required) and extrinsic methods (increased sensitivity due to experimental evidences) were applied for the annotations of the high-quality reference genomes *Arabidopsis (137)*, *Brachypodium (42)*, sorghum *(40)* or rice *(45)*. However, assessments of both approaches have demonstrated that *ab initio* gene finders are highly susceptible to an appropriate selection of a training data set and, in particular, rely on a high quality (at best complete) reference genome sequence *(299,300)*. Many gene models obtained by *ab initio* methods have been shown to be gene fragments or false positive predictions *(301)*. Therefore, additional computational methods and experimental evidences are required to filter incorrect gene predictions or potential pseudogenes *(302)*. For bread wheat both, technical and biological factors, complicate *ab inito* gene prediction on basis on the CSS assembly. Short contigs with a L50 length ranging between 515 bp (3DS) and 4,297 bp (6DS) for individual chromosome arms (Section 2.2) and incomplete assembly or fragmentation of genes on multiple contigs would substantially hamper *de novo* annotation. Additionally, the high abundance of gene fragments and pseudogenes *(199)* may frequently result in false-positive gene predictions. Accompanied by the availability of high-quality reference genomes, decreasing costs and increasing sequencing depth of high-throughput mRNA sequencing enable to generate large-scale transcriptome resources *(142)*. Such data sets open new possibilities for homology-based annotation approaches, which have also been shown to result in the most accurate gene predictions *(299,300)*.

Therefore, in frame of the IWGSC, I implemented a semi-automated extrinsic gene prediction pipeline, which combined different external evidences, including annotated proteins of the closely related reference grasses barley, *Brachypodium*, rice and sorghum, more than 17,000 publicly available wheat fl-cDNAs and a multi-organ wheat RNA-seq data set. Briefly, this gene annotation pipeline consisted of three subsequent steps: First, a set of ”reference-based” gene and transcript structures were generated by merging spliced-alignments of protein sequences of reference genomes and wheat transcriptome sequences (fl-cDNA sequences and *de novo* transcriptome assemblies). Secondly, the initially obtained structures were refined utilizing wheat RNA-seq short reads. Third, the predicted genes were classified in ”high-confidence” (HC) categories, including protein-coding/functional genes, and ”low-confidence” (LC) categories, including nTARs, highly degenerated genes, pseudogenes and gene fragments, based on sequence homology and coverage of available reference plant protein sequence data set.

All subsequent steps and analysis were performed on the repeat-masked version of the CSS assembly (Section 2.2). By using the original, non-masked version of the genome assembly, repetitive and low-complexity sequences would seed spurious alignments, which most likely constitute adverse evidences for the gene annotation *(303)*. Furthermore, the repeat masking reduced the entire search space by 86% and, consequently, decreased the computational time and memory requirements.

4.1.1 Reference-based gene structure prediction

To guide the later assembly of transcript structures by using short RNA-seq reads (Section 4.1.2) and, in particular, to identify non- or low-expressed genes, potential loci were first annotated based on alignments of protein sequences from related grass genomes and from peptide translations of wheat fl-cDNAs and a comprehensive *de novo* transcriptome assembly (Fig. 4.1). Protein-coding wheat fl-cDNAs and assembled transcripts, which were not represented in the CSS assembly, were also identified and completed the set of structurally annotated wheat genes.

De novo assembly of the wheat transcriptome

A comprehensive *de novo* wheat transcriptome assembly was generated independently of the CSS assembly, utilizing Illumina RNA-seq short reads sampled from five different organs (leaf, root, grain, stem and spike) (Section 2.2). Therefore, I pooled the raw RNA-seq reads, resulting in a comprehensive collection of 615 mio reads (62 Gb), and assembled this data set with the *de Bruijn*-graph assembler Trinity *(304)* (release 2012-06-08) and default parameters. The reads clustered into a total of 389,276 contigs (267 Mb) with a mean length of 687 bp and a L50 length of 1.1 kbp (Table 4.1). Utilizing homology supported selection against a combined data set including *Brachypodium (42)*, rice *(45)*, sorghum *(40)*, maize *(41)* and *Arabidopsis (137)* protein sequences the most reliable open reading frames (ORF) were predicted applying the OrfPredictor software *(222)*.

A total of 128,549 contigs were aligned to 94% of public available wheat EST sequences *(203)* (version 1.19 stringent) by using BLASTN *(227)* with an *E* value threshold of 10^{-5}, which

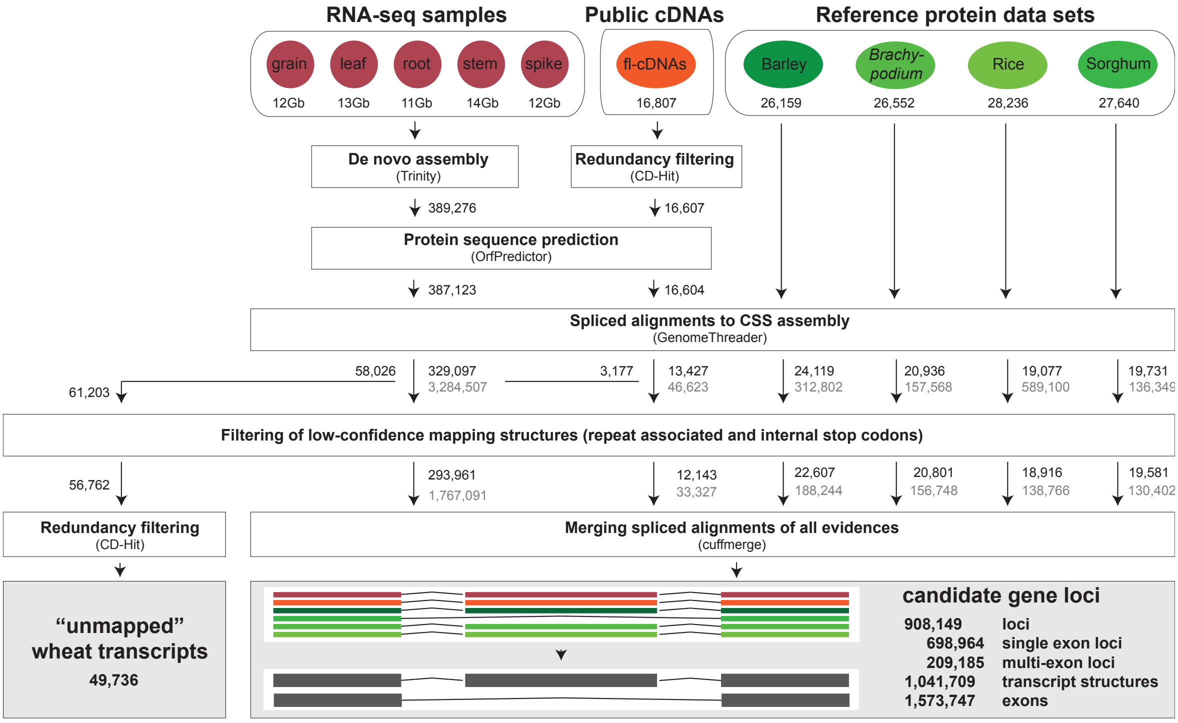


Fig. 4.1. Workflow for the reference-based identification of potential gene structures and wheat transcripts not represented in the CSS assembly.
External protein sequences were spliced aligned against the CSS assembly and stringently filtered for protein-coding potential. Alignments leading to truncated translations of the respective query proteins caused by internal stop codons were removed. A non-redundant structure data set was generated by clustering of structures of different evidences sharing same intron boundaries. Additionally, wheat transcripts were identified, which could not be aligned to the CSS assembly. Black and grey numbers count aligned cDNAs/proteins (RNA-seq assemblies, wheat fl-cDNAs and protein sequences of reference grass genomes, respectively) and distinct GenomeThreader alignments, respectively.

Table 4.1. Assembly statistics of the *de novo* assembly of wheat RNA-seq reads obtained for five organs.

Number of assembled contigs	389,276
Total assembled sequence	267,459,986 bp
Minimum / maximum contig length	201 bp / 31,162 bp
Mean contig length	687 bp
N50 / N90 contig length	1,106 bp / 272 bp
GC content	47.72%
Number of contigs with predicted ORF	387,123
Assembled contigs matching HarvESTs (v1.19)	128,549 (33%)
Matched HarvESTs (v1.19) by assembled contigs	85,618 of 90,786 (94%)

suggested an almost complete representation of the wheat transcriptome. However, the mean length of obtained wheat sequences (approximately 700 bp) was substantially reduced as compared to *Brachypodium* transcripts (approximately 3,000 bp) *(42)*, which indicated highly fragmented transcript sequences. Furthermore, the large number of assembled sequences also indicated that the data set included a significant portion of non-protein-coding transcripts requiring further filtering to distinguish between protein-coding and other transcripts.

Reference-based gene structure prediction

Next, the predicted proteins of wheat cDNA sequences [publicly available wheat fl-cDNAs *(204)* and the *de novo* assembly] and the protein sequences of barley, *Brachypodium*, rice and sorghum were aligned against the CSS assembly with GenomeThreader *(305)* (version 1.5.1 with parameters: –exondistri –refseqcovdistri –prseedlength 7 –species rice –gcmincoverage 0 -force) (Fig. 4.1). The obtained alignments were stringently post-processed to eliminate false positive predictions, which were most likely caused by repetitive mechanisms generating gene fragments or pseudogenes *(199)*. GenomeThreader-alignments were discarded with less than 70% coverage of the query protein or that result in protein translations interrupted by a stop codon, the most radical change leading to inactivated proteins. Furthermore, loci related to repetitive sequence elements were identified and removed by screening the human readable descriptions of the reference protein data sets for the terms ”retrotransposon”, ”transposon”, ”helicase” and ”integrase” as well as by comparing the predicted wheat transcript sequences against the ”Triticeae Repeat Sequence Database”[(1)] with BLASTN *(227)* ($E \leq 10^{-5}$). Finally, a non-redundant set of transcript structures was built by merging the filtered alignments for each reference data set with cuffmerge *(306)* (version 2.0.2).

More than 80% of the wheat transcriptome sequences [329,097 *de novo* assembled transcripts (85%) and 13,427 public available wheat fl-cDNAs *(204)* (81%)] and, corresponding to the evolutionary distances to wheat, between 68% (rice) to 92% (barley) of the reference grass proteins were aligned against the CSS assembly (Fig. 4.1). The reference-based gene structure data set included a total of 908,149 potential gene loci with 1,041,709 distinct transcript structures and 1,573,747 exons. Additionally, 61,203 wheat fl-cDNAs and transcript assemblies were identified, which could not be aligned to the wheat reference genome sequence (15%). Redundant sequences were removed by nucleotide sequence clustering utilizing Cd-Hit *(234)* with 98% nucleotide identity, which resulted in 49,736 wheat transcripts (12%).

4.1.2 Identification of tissue-specific transcript structures

Alternative splicing has been shown to be highly specific for individual tissues or cell types and might be regulated differentially under changing environmental influences *(184,191,307)*. Therefore, I performed reference-guided transcript assemblies of the RNA-seq reads individually for

[(1)] The Triticeae Repeat Sequence Database was downloaded from http://wheat.pw.usda.gov/ITMI/Repeats; version 08/22/2012 each of the five sampled wheat organs (Fig. 4.2). To define the exact alignments of reads across the genome and to avoid wrong mapping of reads to their homoeologous counterparts, an iterative alignment procedure was applied using different stringency levels. Then, transcripts structures were assembled based on the coordinates of these read alignments and supported by the previously defined reference-based gene loci. These provided particularly useful, additional evidence for the definition of low-expressed transcripts or genes absent in the RNA-seq data set.

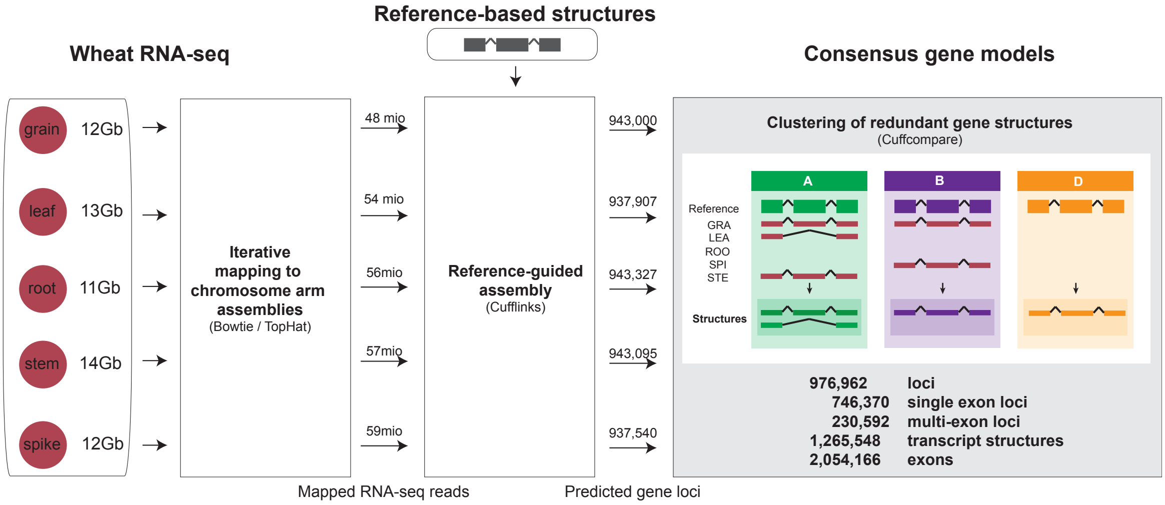


Fig. 4.2. Workflow for RNA-seq based gene prediction and detection of tissue-specific transcripts variants.
Consensus genes and transcript models were created based on reference-based structures and individual transcript assemblies for wheat RNA-seq data of five tissues. See text for a detailed description of individual steps.

Iterative alignment of RNA-seq reads against the genomic reference assembly

For each sampled wheat tissue, the generated RNA-seq reads were aligned in three successive runs against the repeat-masked CSS assembly using Bowtie2 *(174)* (version 2.0.0-beta6) and TopHat *(186)* (version 2.0.3) allowing no, one and two read mismatches, respectively (parameters: read-mismatches 0/1/2 segment-mismatches 1 max-multihits 100). After each iteration I removed reads mapping with best quality score to two or more locations in order to eliminate those for which the genome location and homoeologous genomic origin could not be unambiguously determined. The remaining alignments of uniquely mapped RNA-seq reads were iteratively accepted beginning with the most stringent alignment.

Between 25% to 31% of reads were unambiguously aligned to the CSS assembly considering only perfect alignments of RNA-seq reads and 10% to 17% matched multiple genomic locations (Fig. 4.3a). Both proportions increase up to 49% of unique mapped reads and 27% of multiple mapped reads allowing maximum two mismatches in the read alignments. Overall, the proportion of aligned reads (55% to 71%) was in line with other analysis *(191)*. The iterative alignment strategy identified a total 274 mio high stringency read alignments against the CSS

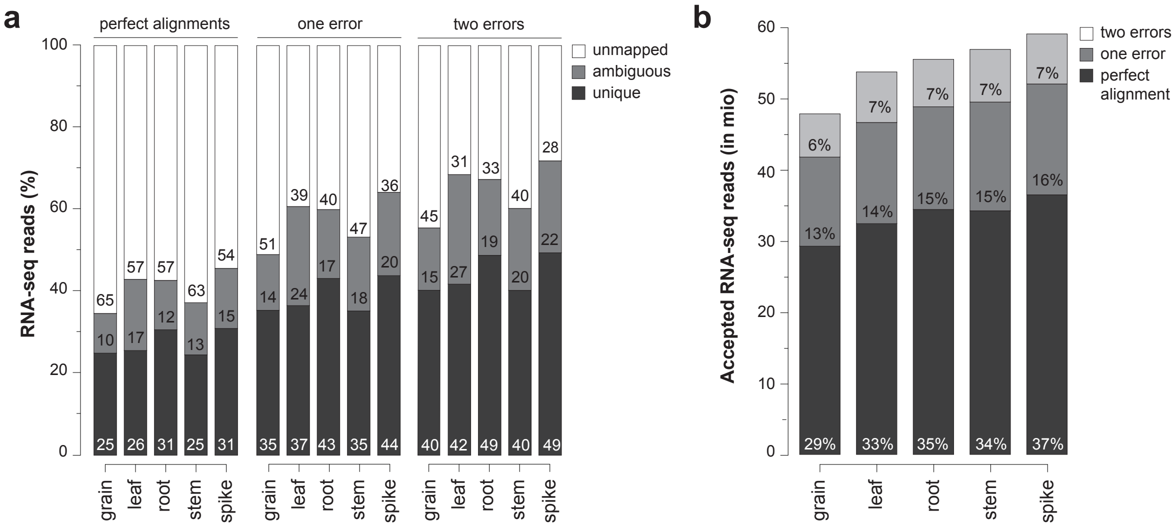


Fig. 4.3. Alignment statistics for RNA-seq reads against the wheat CSS genome assembly.
a, Distribution of unique, ambiguous mapped and unmapped RNA-seq reads for each sample with different alignment stringency levels. **b,** Number of accepted RNA-seq read alignments for each sample using the iterative alignment strategy.

assembly, of which the majority were contributed by perfectly aligned reads (Fig. 4.3b).

Transcript reconstruction and consensus gene modelling

Cufflinks *(306)* (version 2.0.1) was applied for each tissue to assemble the mapped RNA-seq reads into transcript structures utilizing the previously identified homology-based transcript structures as reference annotation (parameter -g). Nucleotide sequences were extracted from the CSS assembly, the most reliable open reading frame determined with OrfPredictor *(222)* and strand direction of individual structures re-defined in accordance to the predicted protein sequence. Finally, all transcript models were clustered based on identical intron boundaries, which were identified in multiple tissues, by using cuffcompare *(306)* (version 2.0.1).

This procedure identified a total of 976,962 loci including 1,265,548 alternative transcripts (i.e. splicing variants) (Table 4.2). Between 29,391 (spike) and 34,851 (grain) novel loci were predicted for individual wheat tissues from RNA-seq short reads alone and not present in the

Table 4.2. Exon, transcript and gene structure prediction statistics for the reference-based annotation, the gain of information with RNA-seq data and the consensus structure set.

	Ref.[(a)]	Novel predictions based on aligned RNA-seq reads					Cons.[(b)]
		grain	leaf	root	stem	spike	
Loci	908,149	34,851	29,758	35,178	34,946	29,391	976,962
Transcripts	1,041,709	91,198	82,276	87,388	92,018	87,792	1,265,548
Exons	1,573,747	280,937	266,337	269,864	281,422	278,951	2,054,166

[(a)] Reference-based gene structure prediction (Section 4.1.1).
[(b)] Consensus gene set determined by clustering transcript structures based on common intron splice sites.

previously defined reference-based structures. In combination, the RNA-seq predictions of the five tissues contributed 68,813 (7%) gene loci to the consensus annotation. Notably, substantial larger relative fractions of additional distinct exons (23%) and alternative splicing variants (18%) were identified. These findings underlined that RNA-seq technology is a valuable resource for the discovery of previously unknown alternatively spliced genes *(184,307,308)*.

4.1.3 Confidence classification of wheat gene predictions

With more than 900,000 predicted loci in the CSS assembly, the consensus set exceeded approximately ten-fold previous estimates of the hexaploid wheat gene repertoire reported in this thesis (Section 3.2.3) and other studies *(43,44,63,241)*. However, this significant increase was triggered by technical and biological factors. Especially for genomes with highly repetitive sequences, NGS-based assemblies are limited in contig length and some gene loci might not be assembled in full-length or split onto multiple contigs. Moreover, besides the high abundance of gene fragments and pseudogenes, which increase substantially gene estimates *(199,241)* (Section 3.3), usage of RNA-seq data evokes the discovery of nTARs in addition to expressed protein-coding genes *(191,309)*. Therefore, I implemented a multi-step bioinformatic pipeline for the further post-processing and confidence classification of the predicted proteins (Fig. 4.4). Based on peptide-homology analysis against high-quality reference plant protein sequences, which served as templates for full-length genes, fragmented wheat loci were identified and the predicted proteins grouped into high-confidence categories (functional, protein-coding genes) as well as low-confidence categories (highly diverged genes, pseudogenes, non-protein-coding transcribed regions).

Identification of template reference genes

First, for each predicted wheat transcript I selected one ”template” peptide from high-quality plant reference gene annotations, which covered comprehensively the flowering plants (angiosperms) and include monocot as well as dicot species (Table 4.3). These template peptides were further used to assess the protein-coding potential of the wheat transcripts, locus fragmentation or pseudogenization. I aligned the wheat sequences against each reference gene set [BLASTP *(227)* ($E \leq 10^{-5}$)] and identified the best scoring alignment for each search. In case of multiple matched reference databases the most suitable template protein was selected on basis of an alignment identity score, which adjusted for different evolutionary distances between bread wheat and the respective reference species (s) by multiplication of the obtained protein alignment identities (i) with a correction factor (C). The corrected alignment scores of a wheat transcript (t) against the reference plant databases (S) were ordered and the reference protein with highest value was selected as template T:

$$T(t) = \max_{s \in \mathrm{S}} \; [\; C(s) \; i(t,s) \;] \tag{4.1}$$

The species-specific correction factors were defined by the sequence similarity distribution be-

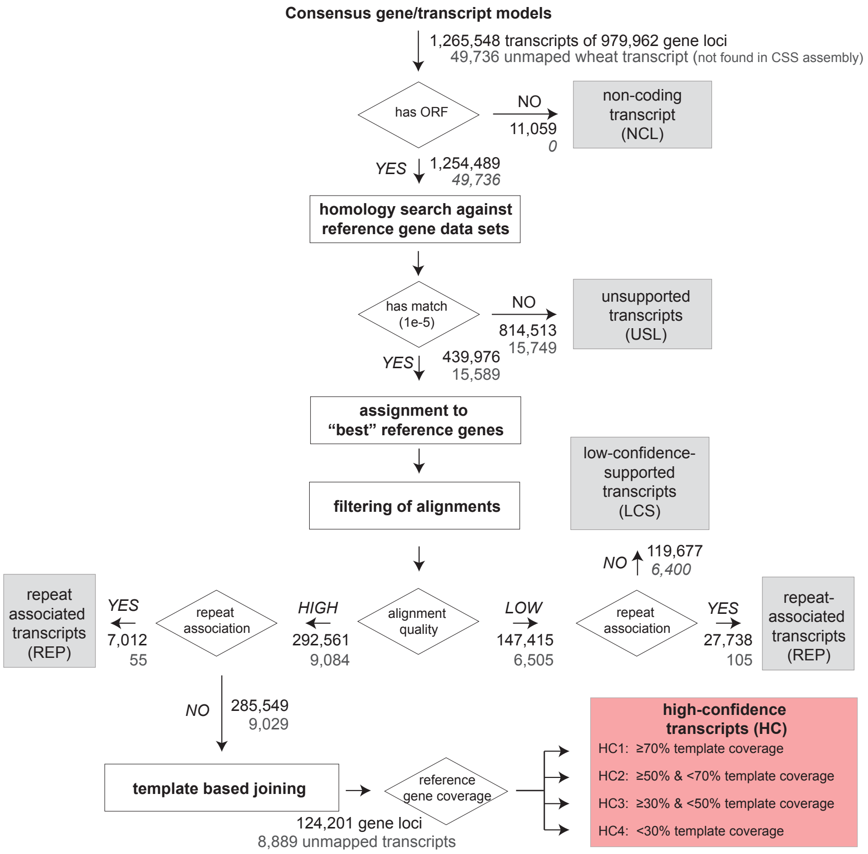

Consensus gene/transcript models
1,265,548 transcripts of 979,962 gene loci
49,736 unmaped wheat transcript (not found in CSS assembly)
has ORF
NO
11,059
0
non-coding transcript (NCL)
YES
1,254,489
49,736
homology search against reference gene data sets
has match (1e-5)
NO
814,513
15,749
unsupported transcripts (USL)
YES
439,976
15,589
assignment to "best" reference genes
filtering of alignments
low-confidence-supported transcripts (LCS)
NO
119,677
6,400
repeat associated transcripts (REP)
YES
7,012
55
repeat association
HIGH
292,561
9,084
alignment quality
LOW
147,415
6,505
repeat association
YES
27,738
105
repeat-associated transcripts (REP)
NO
285,549
9,029
template based joining
reference gene coverage
124,201 gene loci
8,889 unmapped transcripts
high-confidence transcripts (HC)
HC1: ≥70% template coverage
HC2: ≥50% & <70% template coverage
HC3: ≥30% & <50% template coverage
HC4: <30% template coverage

Fig. 4.4. Classification of the consensus gene set into high- and low-confidence gene categories. The predicted wheat gene set was partitioned into high-confidence classes (red box) and low-confidence classes (grey boxes) on basis of conserved sequence homology against public available plant proteins. A detailed description of individual steps is given in the main text. Number of transcripts that were detected in the CSS assembly are shown in black. Number of unmapped wheat transcripts are shown in grey.

tween public available wheat fl-cDNAs ($f \in F$) *(204)*, which were used as ”gold standard” representation of wheat proteins, and each reference species ($s \in S$):

$$C(s) = \frac{100}{\text{mean}_{f \in F}\ i(f,s)} \tag{4.2}$$

Confidence class assignment for predicted wheat gene loci

This work distinguished between HC wheat genes, which were most likely protein-coding genes, and LC wheat genes, which were non-functional or non-protein-coding transcripts. Therefore, I defined a minimum alignment identity cut-off, which represented the expected protein identity for functional wheat genes, by subtracting one standard deviation from the mean alignment identity of wheat fl-cDNAs against each reference proteome [first-best BLASTP *(227)* alignment with *E* value $\leq 10^{-5}$] (Table 4.3). Gene predictions without any match against the reference plant protein sequences were classified as ”unsupported loci” (USL). Furthermore, I screened the

Table 4.3. Reference proteome data sets and parameters used for identification of high- and low-confidence wheat genes.

Set	Species	Description	Σ[a]	$\oslash$[b]	$\geq$[c]	$\mathbf{C}$[d]
1	Wheat	Peptide predictions for a publicly available wheat fl-cDNAs *(204)*	16,604	-	90%	1.00
2	Barley	Barley high-confidence gene set *(191)*	26,159	88% ± 14%	74%	1.12
3	*Brachypodium*, rice, sorghum, barley (fl-cDNAs)	Conserved grass orthologous gene representatives which were defined by orthologous clustering *(310)*	20,401	81% ± 16%	65%	1.23
4	Maize, *Arabidopsis*	Ensemble gene set that incorporates the proteomes of the more distant plant species maize *(41)* and *Arabidopsis thaliana (137)*	95,876	74% ± 15%	59%	1.35

[a] Number of sequences

[b] Mean BLASTP *(227)* alignment identity and standard deviation to wheat fl-cDNAs.

[c] Minimum alignment identity threshold used for confidence assignment.

[d] Correction factor applied for selection of reference template genes.

selected template genes of all transcripts for repeat associated descriptions (”retrotransposon”, ”transposon”, ”helicase” or ”integrase”) and collected those genes in a separate class of ”repeat-associated” (REP) loci. All remaining transcripts, which had reduced protein sequence homology (i.e. alignment identity <cut-off), were grouped in the ”low-confidence supported” (LCS) category, while transcripts with the expected protein sequence homology (i.e. alignment identity ≥cut-off) were grouped in the HC gene set.

Overall, an alignment against at least one reference protein was identified for 439,976 out of 1,254,489 (35%) wheat transcripts, which were predicted in the CSS assembly and 15,589 out of 49,736 (31%) unmapped cDNA sequences (Table 4.4). Two thirds of these transcripts (292,561 and 9,084, respectively) satisfied the defined alignment identity cut-off for HC transcripts. Barley proteins were selected as template for the majority of wheat transcripts (174,606 and 2,885, respectively) reflecting the close evolutionary distance of the two genomes (Section 1.2.2). Whilst less then three percent of wheat transcripts with high alignment identity above the defined cut-off were associated to repeat elements, almost 20% of the annotated transcripts with low alignment identity fell into the REP confidence class. Thereby, the repeat-association was far less pronounced for unmapped wheat transcripts.

Template based joining of fragmented gene loci

The defined template reference genes were also utilized to detect gene loci, which have been split on two or more contigs in the CSS assembly, or unmapped wheat transcripts, which were only partially assembled *de novo*. Therefore, I computed protein alignments between wheat transcripts and the cognate template peptides (Fig. 4.5a) and determined the coverage of the reference genes (Fig. 4.5b). If two or more splicing variants of a gene were aligned, I selected the transcript with highest coverage as representative and discarded all other alignments. Then, I constructed a directed graph for each template protein representing all accepted alignments with respect to the

Table 4.4. Alignment statistics for comparison of predicted wheat transcripts positioned within the CSS assembly against the reference proteome data sets.

	Set 1[a]	**Set 2**[a]	**Set 3**[a]	**Set 4**[a]	**Total**
Predicted wheat transcripts located in the CSS assembly					
Aligned transcripts	323,992	399,430	404,488	332,538	439,976
Low sequence similarity alignments					
Aligned transcripts	221,457	162,857	186,529	121,362	147,415
Transcripts with template[b]	16,980	49,935	61,104	19,396	147,415
Selected references	2,474	3,712	2,795	3,377	12,358
Transcripts associated to repeat[c]	-	19,477	7,504	757	27,738
Gene loci with template(s)	12,156	43,046	54,843	15,238	122,382
High sequence similarity alignments					
Aligned transcripts	102,535	236,573	217,959	211,176	292,561
Transcripts with template[b]	36,051	174,606	48,895	33,009	292,561
Selected references	10,855	20,185	7,712	7,841	46,593
Transcripts associated to repeat[c]	-	6,002	963	47	7,012
Gene loci with template(s)	19,794	78,471	30,885	20,969	132,554
Predicted wheat transcripts not found in the CSS assembly					
Aligned transcripts	12,680	13,694	13,753	13,213	15,590
Low sequence similarity alignments					
Aligned transcripts	8,419	7,952	7,706	6,680	6,505
Transcripts with template[b]	367	386	948	4,804	6,505
Selected references	246	309	704	3,559	4,818
Transcripts associated to repeat[c]	-	39	33	33	105
High sequence similarity alignments					
Aligned transcripts	4,261	5,742	6,047	6,533	9,084
Transcripts with template[b]	3,140	2,885	1,052	2,007	9,084
Selected references	2,955	2,303	793	1,556	7,607
Transcripts associated to repeat[c]	-	37	8	10	55

[a] For definition of reference protein resources see Table 4.3.

[b] Transcripts with selected templates from respective reference proteome.

[c] Repeat information was not available for the wheat fl-cDNA reference data set *(204)*.

exact start and end alignment coordinates (Fig. 4.5c). Aligned genes were represented as nodes weighted by the observed template coverages. Potentially neighbouring genes were connected by edges, if the two loci (i) aligned to the same template gene, (ii) were annotated on different contigs of the same chromosome arm, (iii) had non-overlapping alignments along the template protein and (iv) were separated by gaps with less than 30 amino acids length. I directed the edges based on the alignment coordinates along the template proteins, starting from the amino terminus to the carboxyl terminus. Fourth, the entire graph was split into connected components, for which I computed all possible paths that include all nodes at least once using a top-down search algorithm (Fig. 4.5d). These paths were scored by the sum of the node weights to determine the total coverage of the template gene obtained by the combination of the aligned transcripts. Then, I selected the path with maximum score as most-likely combination of split genes, marked the combined loci as fragmented and defined the gene with maximum template coverage as representative for the path (Fig. 4.5e). Finally, used nodes and edges were removed from the graph

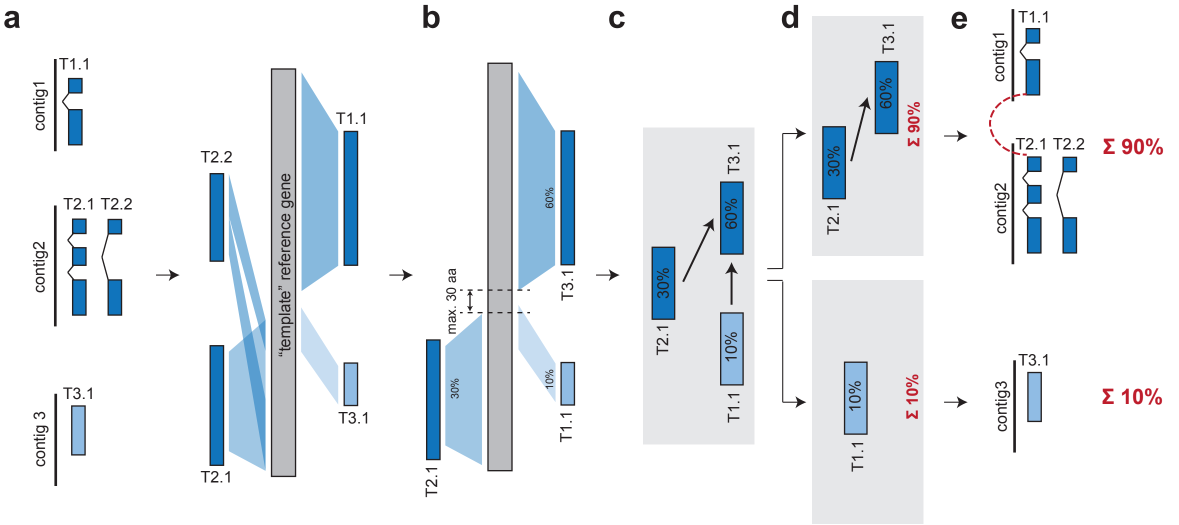


Fig. 4.5. Pipeline for joining of fragmented gene loci based on the alignments to the associated template reference peptide.
See main text for detailed description of individual steps.

and the identification of maximum scoring routes repeated until no nodes remained.

A total of 285,549 non-repetitive wheat transcripts, which were positioned in the CSS assembly and corresponded to 127,093 gene loci, as well as 9,029 unmapped wheat transcripts had high-confidence alignments (alignment identity ≥cut-off) against the reference proteins (Fig. 4.4). These were subjected to the template-based joining algorithm (Table 4.5), whereupon 40,387 and 7,560 template reference genes were selected. By using the above-defined criteria 3,929 and 162 edges between two gene loci form an initial set of 92,257 and 7,560 connected components. The graph traversal determined 5,643 and 273 mutually completing HC gene loci and transcripts, respectively. This resulted in a final HC protein set of 133,090 genes of which 124,201 were positioned within the CSS assembly (93%) and 8,889 were represented by unmapped wheat transcripts (7%).

Assignment of confidence classes to predicted wheat gene loci

Finally, the predicted genes were grouped into different confidence classes based on sequence homology and protein coverage of the previously selected template reference genes (Fig. 4.4). For alternatively spliced loci the confidence class of the best supported transcript was passed to the entire gene locus, i.e. to all splice variants. A total of 133,090 loci showed high-similarity homology with related grass proteins and, thus, were classified as high-confidence, protein-coding gene predictions (Fig. 4.6 and Table 4.6). These were further subdivided into four levels based on the coverage of the template reference gene (Fig. 4.6a). 59,426 gene loci covered at least 70% of the reference template gene and were classified as HC level 1 (HC1) genes (Fig. 4.6b). This group represented the most reliable confidence class level and included (almost) full-length wheat gene predictions. Three further HC confidence levels were defined, of which the protein

Table 4.5. Statistics for template based joining of mutually completing gene loci.

	Transcripts found in the CSS assembly	Transcripts not found in CSS assembly
High-similarity alignments to reference proteomes		
Transcripts	285,549	9,029
Gene loci	127,093	9,029
Template reference gene set		
Set 1 (wheat fl-cDNAs)	9,471	2,955
Set 2 (barley)	18,893	2,272
Set 3 (*Brachypodium*, rice, sorghum)	6,030	785
Set 4 (maize, *Arabidopsis*)	5,993	1,548
Total	40,387	7,560
Template reference gene set		
Created nodes (loci)	127,093	9,029
Created edges (connection of loci)	3,929	162
Connected components	92,257	7,560
Loci joined by traversal of graph structures	5,643	273
Final number of gene loci after traversal of graph structures	124,201	8,889

coding confidence decreases with decreasing template reference gene coverage. While levels HC2 (≥50% and <70% coverage) and HC3 (≥30% and <50% coverage) represented medium confidence class levels, level HC4 was the least reliable confidence level, which loci spanned only a small proportion of a reference protein sequence (<30% coverage). With decreasing confidence levels a trend for higher divergence of the corresponding transcripts was observed (Fig. 4.6c). Loci with detectable but substantially reduced sequence similarity to the reference proteins were defined as low-confidence supported gene loci (95,398 loci). The remainder were classi-

Table 4.6. Overview of the confidence classification for predicted wheat gene loci.

Confidence class		Description	Found in the CSS assembly	Not found in the CSS assembly
High-confidence (HC) genes: High sequence similarity alignments against plant reference protein sequences				
HC1	Level 1 HC	Reference coverage $\geq$70%	55,249	4,177
HC2	Level 2 HC	Reference coverage $\geq$50% and $<$70%	14,367	662
HC3	Level 3 HC	Reference coverage $\geq$30% and $<$50%	15,475	1,053
HC4	Level 4 HC	Reference coverage $<$30%	39,110	2,997
Σ			124,201	8,889
Low-confidence loci: Reduced or no sequence similarity against plant reference protein sequences				
REP	repeat-associated loci	Aligned to a reference gene which was classified as repeat based on its gene description	32,412	160
LCS	low-confidence-supported	Aligned to a reference gene with low sequence similarity	88,998	6,400
USL	unsupported loci	No match against reference proteomes, but ORF predicted	718,048	34,147
NCL	non-coding	No ORF predicted by OrfPredictor	10,411	0
Σ			849,859	40,707

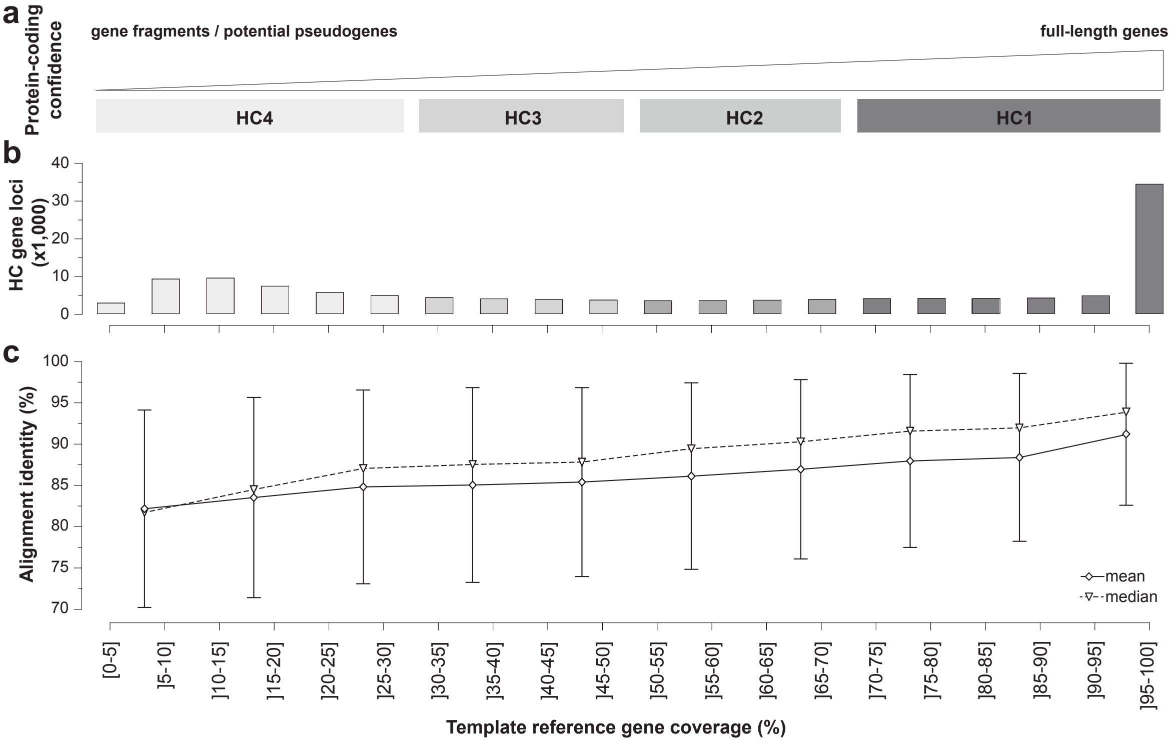


Fig. 4.6. Template reference gene coverage and sequence similarity of high-confidence gene loci. **a,** Based on template reference gene coverage the high-confidence wheat gene loci were categorized into four confidence classes, HC1 to HC4. **b,** Frequency distribution of predicted HC genes. **c,** Mean and median alignment identity of predicted wheat genes against the selected template protein.

fied as repeat-associated genes (32,572), non-coding loci (10,411) or unsupported loci without homology to any reference plant protein sequence (752,195).

4.2 Evaluation of the wheat gene annotation

4.2.1 Influences of sequencing depth and assembly quality on the HC gene set

The sequencing depth varied considerable for individual chromosome arms and ranged from 28-fold for the short arm of chromosome 7A up to 242-fold for the short arm of chromosome 4A (Fig. 4.7a). Also, the assembly quality, here defined by the L50 contig length of the repeat-masked CSS assembly, differed largely between 1.7 kbp (3DL) and 8.9 kbp (6DS) (Fig. 4.7b). However, no significant correlation between these technical criteria was observed [Pearson's correlation coefficient (R^2) of 0.04]. With respect to the gene annotation I further tested each chromosome arm for dependency between sequencing depth, assembly quality and the absolute number of predicted HC gene loci as well as the proportion of HC1 genes, respectively (Fig. 4.7c).

Neither technical measurements showed a significant influence on the total number of predicted high-confidence genes ($R^2 < 0.02$). This indicated that the general detection of genes was not substantially influenced by the differences in sequencing depth and contig L50 length between single chromosome arms suggesting saturation and completeness for all chromosome arm gene sets. In contrast, the relative number of HC1 genes predicted for individual chromosome arms had weak correlations with sequencing-depth (R^2 of 0.31) and assembly L50 (R^2 of 0.46). Although there was no direct correlation between the two technical measurement themselves, the increased proportion of predicted HC1 suggested that gene loci were more likely to be assembled in full-length for chromosome arms with improved quality measures. This indicated that increased sequencing depth has been beneficial in particular for the complete assembly of gene loci and reduced the fragmentation of genes into multiple contigs.

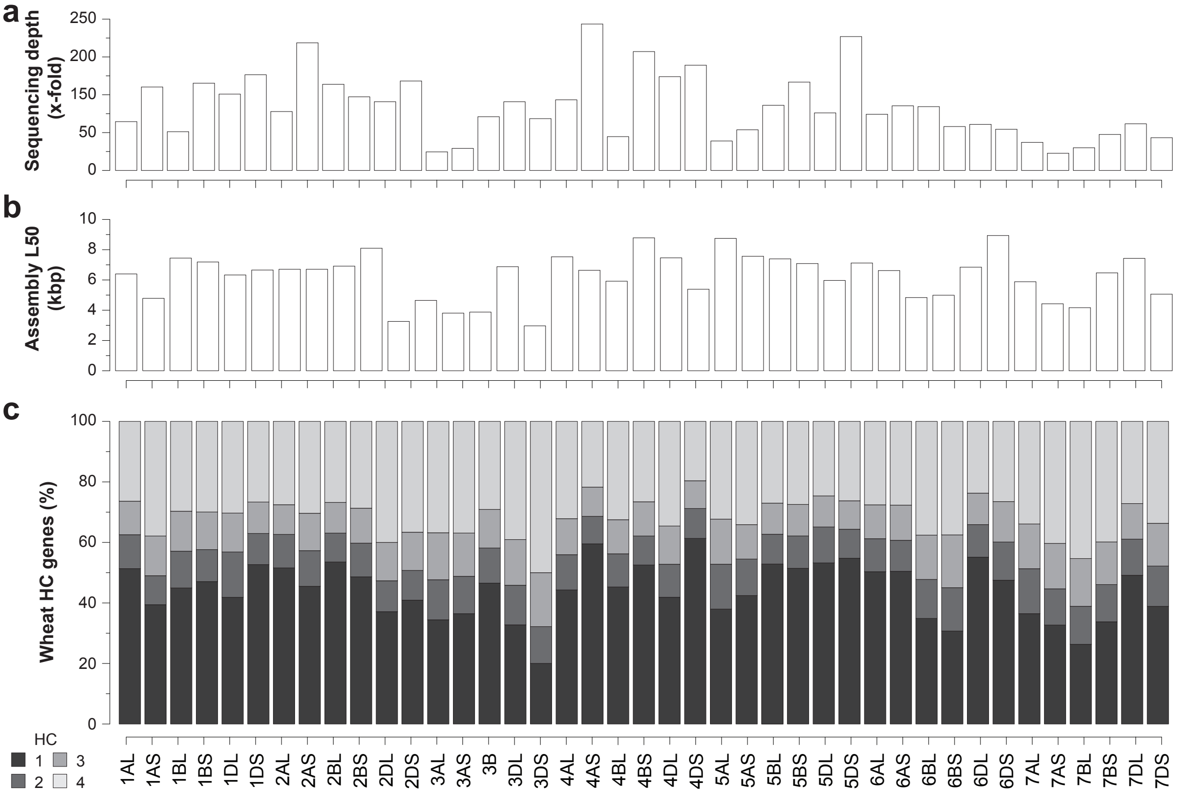


Fig. 4.7. Sequencing depth, assembly quality and distribution of high-confidence gene loci for each chromosome arm.

a, Sequencing depth, **b,** assembly quality represented by the contig L50 length of the repeat masked CSS assembly and **c,** the proportion of the high-confidence gene loci in among confidence class levels for individual chromosome arms.

4.2.2 Completeness of the predicted bread wheat gene set

To further evaluate the completeness of the gene annotation, I compared the predicted wheat transcripts against an independent set of bread wheat EST sequences, which are publically available in the HarvEST database *(203)* (version 1.19 stringent), by using BLASTN *(227)* with a maximum *E* value of 10^{-5}. A total of 206,778 wheat genes annotated in the CSS assembly and 8,148 unmapped wheat transcripts matched 87,389 and 8,148 out of 90,786 HarvESTs, respectively (Fig. 4.8a). In combination, 61% of all high-confidence gene loci (81,274) contributed to the detection of more than 90% of the EST sequences. Ninety-seven percent of the EST sequence were tagged by considering both, high- and low-confidence genes (87,963). The large majority of HarvEST sequences (85%) were already matched by considering HC1 genes only.

With decreasing (protein-coding) reliability of individual confidence classes less annotated genes were aligned against the HarvESTs (Fig. 4.8b). Approximately half of the HC4 and less than 20% of low-confidence gene loci exhibited significant BLAST *(227)* matches and did not significantly contribute to the overall detection of EST sequences. The reduced representation of HC4 and LC gene predictions in a comprehensive public wheat database provided further evidence that, most likely, the majority of these gene calls constituted non-expressed gene fragments and pseudogenes, a conclusion that was supported by the decreased in protein sequence conservation (Section 4.1.3). In summary, these findings indicated that the structurally defined high-confidence gene set was highly representative for the entire gene inventory of the bread wheat genome and that more than 90% of wheat genes have likely been captured in this study.

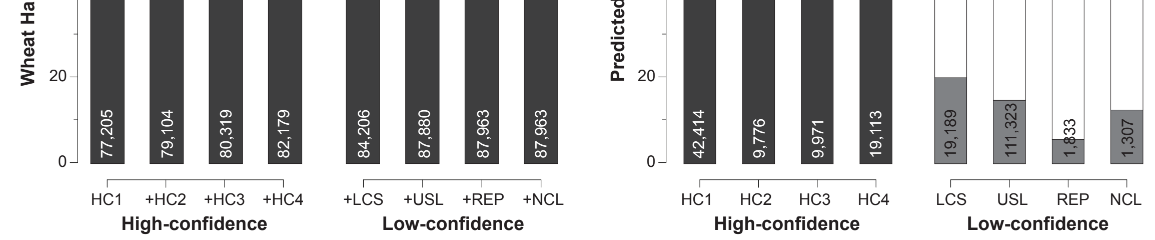


Fig. 4.8. Comparison of bread wheat genes against publicly available wheat EST sequence assemblies of the HarvEST database.
a, Cumulative number of matched (filled proportion of bars) and not matched wheat EST sequences (non-filled proportion of bars) for different confidence classes. Dark and light grey represent the fraction of ESTs, which were matched by HC gene(s) and exclusively by LC gene(s), respectively. **b,** Fraction of annotated wheat genes of which at least one transcript was aligned to publicaly available wheat EST sequences.

4.2.3 Estimation of the bread wheat gene number

The total gene number of the bread wheat genome was estimated on the basis of the level of completion of the CSS assembly and the detection of HC1 genes. Therefore, a ”gold standard” wheat reference set was defined utilizing approximately 17,000 publicly available wheat fl-cDNAs *(204)*. These were first allocated to individual chromosome arms via the ”Chromosome arm Assigner” (CarmA) method, a computational approach originally developed for barley *(191)* and subsequently adapted to distinction of homoeologous relationships and chromosome-of-origin within hexaploid wheat. Briefly, CarmA homology-searches were conducted for the fl-cDNA sequences against the CSS assembly and the most likely chromosomal origin was determined[(2)]. Ninety-one percent (15,300) of the fl-cDNAs were allocated to individual chromosomes, with a relative even distribution between the homoeologous genomes [A genome: 5,023 fl-cDNAs (32.8%), B genome: 5,344 fl-cDNAs (34.9%) and D genome: 4,933 fl-cDNAs (32.2%)].

Subsequently, I evaluated the completeness of the CSS assembly by comparing four different genome data sets obtained for chromosome 3B against 966 wheat fl-cDNAs assigned to this chromosome: (i) the original, non-masked version of the CSS assembly, (ii) the repeat-masked version of the CSS assembly, (iii) the raw Illumina 3B sequence reads and (iv) 3B scaffolds produced in a BAC-based sequencing project *(311)* (Fig. 4.9a). All queries were mapped against the fl-cDNAs using VMATCH *(226)* and alignments were filtered requiring 100% sequence identity spanning at least 50 bp. Furthermore, to elucidate the loss of protein coding information during assembly, I compared the expected coverage of reference proteins by direct alignments of wheat fl-cDNAs with the obtained coverage with respect to the representation of a fl-cDNA in the CSS assembly (Fig. 4.9b). For each fl-cDNA I identified a template reference gene using the same procedure as previously described (Section 4.1.3). Additionally, I identified the best-coverage spliced-alignment of a fl-cDNA against the repeat-masked CSS assembly (Section 4.1.1) and adjusted the direct peptide alignments for the observed representation of the fl-cDNA in the genomic sequences. Reduced coverage of fl-cDNAs and reference grass genes was obtained for alignments via the CSS assembly in comparison to the direct peptide alignments.

On the one hand, the detection and coverage of fl-cDNA sequences varied substantially between different sequence types and sources, whilst only minor differences between the original and repeat-masked versions of the CSS assembly were observed (Fig. 4.9c). Compared to the CSS raw reads and assemblies, on which the fl-cDNA chromosome assignment was based, a total of 101 fl-cDNAs were not detected using the 3B BAC-based scaffolds, which was in line with the reported completeness of this data set *(311)* . Fl-cDNAs were best covered by raw Illumina sequencing reads and 3B BAC-based scaffolds, whereas the coverage of fl-cDNAs by contigs of the CSS assembly decreased 10% compared to the raw reads indicating loss of sequence information during the assembly process. Consequently, 90% of the entire gene space on basis

[(2)] I gratefully acknowledge Heidrun Gundlach for performing the CarmA analysis. Following parameters were used for CarmA assignment of individual fl-cDNAs assignment to chromosome arms: (i) homology search: forward and reverse strand search, minimum hit length 120 bp and perfect matches; (ii) chromosome arm assignment: highest coverage bin, signal noise ratio between the highest and second highest bin $\geq$1.5.

of the CSS assembly could be predicted as at least 70% coverage of the template genes was required for assignment of gene predictions to confidence class HC1. On the other hand, the comparison between direct alignments of wheat fl-cDNAs to high-quality reference plant protein sequences and indirect alignments via the CSS assembly indicated approximately 20% loss of protein-coding sequence in the CSS assembly (Fig. 4.9d). Therefore, requiring at least 70% coverage of the template protein would classify 58% of the gold standard wheat fl-cDNAs as full-length (HC1) genes.

These two estimators, the assembly completeness ($c_{\mathrm{HC1}} = 0.90$) and the detection rate ($d_{\mathrm{HC1}} = 0.58$), were used to compute the estimated gene number G of bread wheat with respect to the 55,249 predicted HC1 gene loci (G_{HC1}) (Table 4.6) as follows:

$$G = \frac{G_{\mathrm{HC1}}}{d_{\mathrm{HC1}}}\, c_{\mathrm{HC1}} = \frac{55,249}{0.58}\, 0.90 = 105,841 \tag{4.3}$$

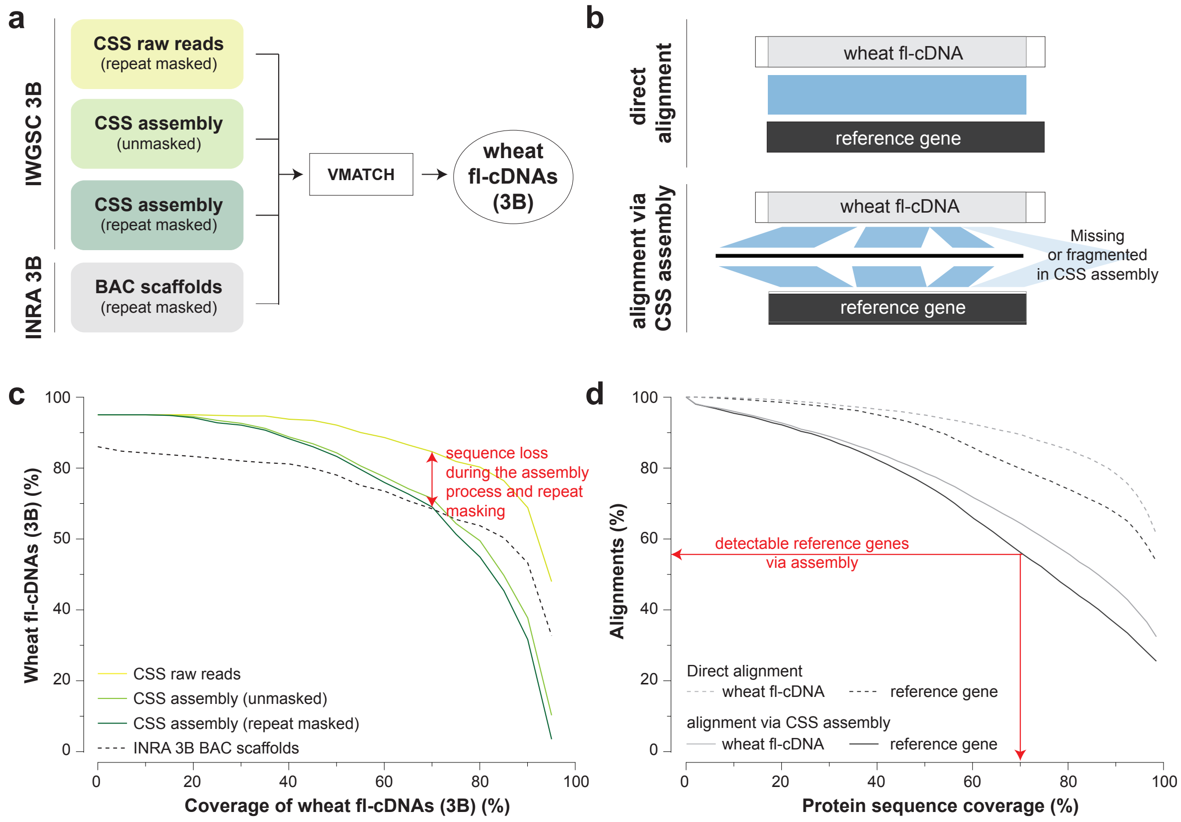


Fig. 4.9. Identification of parameters for estimation of the bread wheat gene content. The bread wheat gene content was estimated based on **a,** completeness of the genomic reference sequences and **b,** the detection rate for HC1 genes in the CSS assembly. **c,** Detection and cumulative coverage of chromosome arm assigned wheat fl-cDNAs by different genomic sequence types and sources for chromosome 3B. **d,** Comparision of the cumulative coverage of template proteins obtained in direct alignments with wheat fl-cDNAs as well as indirect alignments of wheat fl-cDNAs via the CSS assembly.

4.3 Characteristics of bread wheat genes

4.3.1 Structural characteristics of high- and low-confidence wheat genes

As discussed in the previous sections of this chapter, the CSS assembly made it possible to structurally define genes for almost the entire wheat genome. For the first time, this annotation allowed analysing the structural characteristics of wheat genes on a genome-wide level. Besides the ability to encode for functional, protein-coding genes the structural features differed substantially between high- and low-confidence genes and transcripts (Table 4.4).

Locus, transcript and exon length
The mean locus length (including UTR, exon and intron sequences) and mean transcript length (including UTR and exon sequences) were 2.2 kbp and 1.3 kbp for HC loci and substantially longer as for LC loci (0.7 kbp mean locus and transcript lengths). The gene length and transcript length differed also for individual HC confidence levels, ranging from a mean of 3.3 kbp (gene) and 1.6 kbp (transcript) in the HC1 set to 0.9 kbp (gene) and 0.7 kbp (transcript) in the HC4 set. This finding was consistent with the observed template gene coverages for genes of different confidence class levels. Remarkably, almost no variation were found in the median exon sizes among all four HC levels (168 bp to 171 bp) indicating that individual exons of protein-coding genes were most likely assembled complete and fragmentation of genes occurred predominantly in introns. Considering only HC1 genes the observed length were largely in line with predictions in the model grass species *Brachypodium (42)*, rice *(45)* and sorghum *(40)*. In contrast to the reduced locus and transcript sizes, the median exon lengths were considerably increased for low-confidence loci (207 bp to 282 bp).

Exon frequency
Between 17% (HC1) to 52% (HC4) of HC loci were single exon genes. Again, the variances between genes of the four HC levels could be explained by the respective template coverages and locus lengths. However, the observed fractions of single and multi exon genes were largely consistent with observations in other sequenced grass genomes, like *Brachypodium* (21%) *(42)* or barley (25%) *(191)*. Contrarily, approximately three out of four LC loci consisted of a single exon. On average, HC protein-coding transcripts were composed of 5.1 exons, consistent with predictions in *Brachypodium (42)* (5.5 exons per transcript), while LC transcripts consisted in average less exons (2.1 exons per transcript).

Alternative splicing
Accompanied with increased exon frequency, alternative splicing was more prevalent for HC genes compared to LC loci. Two or more alternative transcripts were annotated for approximately half of the HC genes (49%), whilst multiple transcripts were predicted for only 24% of LC genes. This was even more pronounced for HC1 gene loci, of which almost 70% had alternative

Table 4.7. Structural characteristics of high-confidence and low-confidence wheat genes.

High-confidence	**HC1**	**HC2**	**HC3**	**HC4**	Σ
Gene loci	55,249	14,367	15,475	39,110	124,201
Single exon	9,181 (17%)	3,230 (22%)	4,906 (32%)	20,375 (52%)	37,692 (30%)
Multi exon	46,068 (83%)	11,137 (78%)	10,569 (68%)	18,735 (48%)	86,509 (70%)
Alternatively spliced	38,059 (69%)	7,916 (55%)	6,465 (42%)	8,728 (22%)	61,168 (49%)
Mean size (bp)	3,319	2,204	1,608	901	2,216
Median size (bp)	2,747	1,681	1,105	458	1,398
Transcripts	194,624	37,116	31,957	61,450	325,147
Mean[(a)]	3.52	2.5	2.07	1.57	2.62
Median[(a)]	3	2	1	1	21
Maximum[(a)]	46	43	30	27	46
Mean size (bp)	1,626	1,196	983	675	1,334
Median size (bp)	1,422	1,020	794	435	1,112
Distinct exons[(b)]	538,250	94,864	74,630	117,530	825,274
Mean[(c)]	9.74	6.60	4.82	3.01	6.64
Median[(c)]	8	5	3	1	4
Maximum[(c)]	99	85	71	81	99
Mean[(d)]	6.29	4.45	3.52	2.56	5.1
Median[(d)]	5	3	3	2	4
Maximum[(d)]	76	38	39	29	76
Mean size (bp)	321	315	314	281	314
Median size (bp)	168	171	187	186	172
Low-confidence	**LCS**	**REP**	**USL**	**NCL**	Σ
Gene loci	88,998	32,412	718,048	10,411	974,070
Single exon	59,790 (67%)	28,386 (88%)	9,212 (85%)	9,212 (88%)	745,010(76%)
Multi exon	29,208 (33%)	4,026 (12%)	108,118 (15%)	1,199 (12%)	229,060 (24%)
Alternatively spliced	9,798 (11%)	1,210 (4%)	28,484 (4%)	20 (0%)	100,680 (10%)
Mean size (bp)	862	570	423	287	695
Median size (bp)	478	350	273	229	308
Transcripts	113,507	35,285	777,010	10,433	1,261,382
Mean[(a)]	1.28	1.09	1.08	1.00	1.29
Median[(a)]	1	1	1	1	1
Maximum[(a)]	58	20	30	3	58
Mean size (bp)	815	568	390	252	675
Median size (bp)	519	357	271	221	350
Distinct exons[(b)]	192,304	45,886	970,698	11,935	2,046,097
Mean[(c)]	2.16	1.42	1.35	1.15	2.1
Median[(c)]	1	1	1	1	1
Maximum[(c)]	64	58	61	10	99
Mean[(d)]	2.19	1.54	1.34 / 1	1.14	2.39
Median[(d)]	1	1	1	1	1
Maximum[(d)]	28	33	25	8	76
Mean size (bp)	396	391	293	220	313
Median size (bp)	264	282	234	207	222

[(a)] Number of transcripts per locus.

[(b)] Exons of two or more transcripts were counted once if they have identical start and stop positions.

[(c)] Number of exons per locus.

[(d)] Number of exons per transcript.

transcript structures. This remarkable level of alternative splicing, which was generally consistent with recent estimates in *Arabidopsis (184)* and barley *(191)*, and its potential impacts on gene expression regulation will be further investigated and discussed the following section of this chapter.

4.3.2 Genome distribution of protein-coding genes

Overall, 124,201 HC protein-coding genes were structurally defined in the bread wheat genome assembly. Thereby, similar number of genes were obtained for the wheat A genome [40,253 genes (33%)] and D genome [39,425 genes (32%)], while a higher number of genes was detected in the B genome [44,523 genes (35%)] (Fig. 4.10a). This relative distribution was also found only considering genes of an individual confidence class [e.g. A genome: 17,635 HC1 genes (32%), B genome: 20,144 HC1 genes (34%) and D genome: 17,470 HC1 genes (33%)]. Interestingly, the overall gene content distribution was not retained at the chromosomal level. For instance, the gene distribution over homoeologous group 3 chromosomes was 30% for the A genome, 42% for the B genome and 28% for the D genome, whereas the D genome contained the highest proportion of genes for homoeologous group 7 chromosomes (Fig. 4.10b).

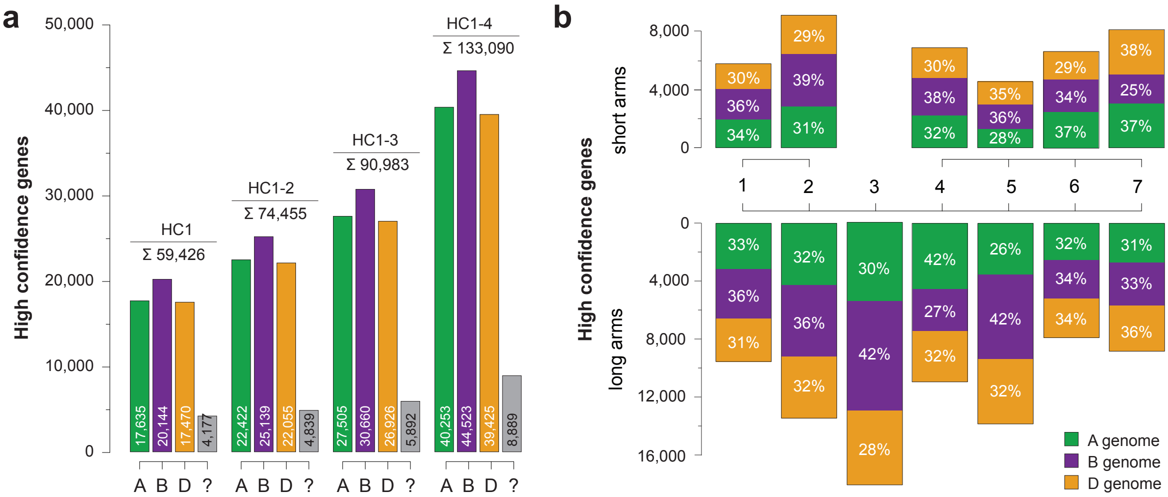


Fig. 4.10. Distribution of high-confidence wheat genes across genomes and chromosomes. **a,** Number of predicted HC wheat genes cumulative for different confidence classes across the A, B and D genome. Numbers for wheat transcripts not found in the CSS assembly are shown by grey bars. **b,** Number of high-confidence genes (HC1-4) for individual chromosome arms or chromosomes (group 3).

The gene density varied up to 2.4-fold between different chromosome arms ranging from 4.4 loci per Mb (5AS) up to 10.4 loci per Mb (2DL) (Fig. 4.11). To investigate the degree of syntenic conservation of individual wheat chromosome arms and *Brachypodium*, rice and sorghum, I further compared the overall gene density against the density observed in the GenomeZipper[3], a

[3] I gratefully acknowledge Mihaela M. Martis, who developed, implemented and performed the GenomeZipper for the bread wheat genome.

synteny-derived approximation of the linear gene order along each chromosome (Section 1.5.1). On average, 53% of the HC genes were located at syntenic positions in the GenomeZipper (Fig. 4.11) on a genome-wide level. The degree of syntenic conservation varied considerably between 34% (6BS) and 67% (5DL) for individual chromosome arms as well as between the A, B and D genomes. The average level of synteny for genes located on the D genome chromosomes (58%) was higher than the average for those on the A chromosomes (51%) and on the B chromosomes (50%). Furthermore, compared to HC genes, LCS genes showed substantially reduced syntenic conservation across all chromosome arms.

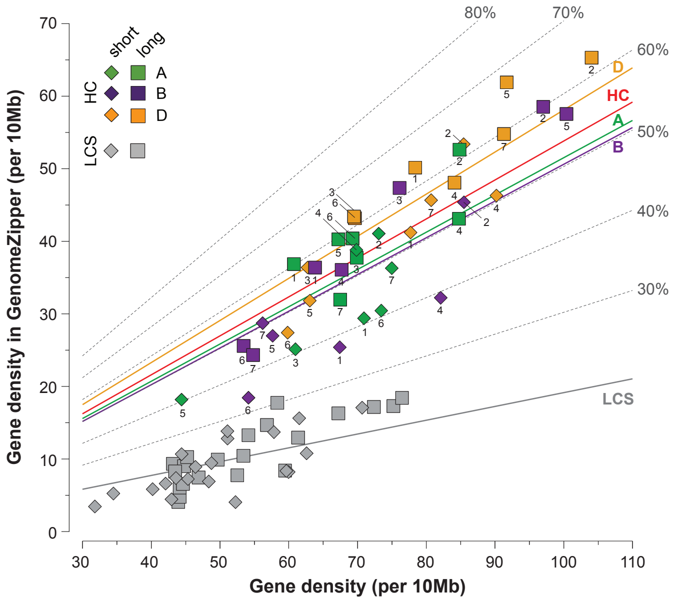


Fig. 4.11. Gene density and syntenic conservation of high-confidence genes and low-confidence supported genes for individual chromosome arms.
Triangles and squares visualize gene density against syntenic conservation for individual short and long chromosome arms (entire chromosome 3B is represented as square). Solid lines show the average syntenic conservation for low-confidence supported (LCS) and high-confidence (HC) genes.

4.3.3 Analysis of homoeologous genes retained in each genome of polyploid bread wheat

Numerous comparative analyses between the bread wheat A, B and D genomes require the identification of homoeologous genes, which were derived from the diploid progenitor genomes and have been retained in hexaploid wheat. In particular, the definition of ”homoeologous gene triplets”, which are formed by genes present in a single-copy in each genome, would permit to investigate, for example, conservation in genome structure, sequence evolution, phylogenetic relationships or homoeolog-specific gene expression regulation. However, current studies aiming at answering these questions have been mostly based on a few (selected) homoeologs due to the lack of comprehensive and suitable genomic resources *(123,312)*. Here, on the basis of the previously unknown gene annotation, a total of 7,228 homoeologous gene triplets were generated[(4)] by using a best-bidirectional hit approach among A-, B- and D-genome encoded HC protein sequences *(313)*. These represented almost twenty percent of the of the entire wheat gene catalogue and incorporated a total of 21,684 (7,228 × 3) genes.

Synteny relationships in homoeologous gene triplets

The large majority of the identified homoelogous gene triplets [6,926 triplets (96%)] consisted of genes that were located on corresponding homoeologous chromosome arms (Fig. 4.12a). Only 302 triplets were formed by unexpected chromosome arm pairings and showed mainly interchange of arm assignments for one member (e.g. 11 triplets were formed by genes of chromosomes 7AS, 7BL and 7DS). Most probably this was caused by contaminations during the chromosome flow-sorting process, which purity has been estimated to be approximately 90% *(70)*. Notably, the findings reflected well the known evolutionary dynamics of chromosomes 4A and 5A, respectively *(63,314,315)* (Fig. 4.14b and the following section). Consistent with the translocation of a chromosomal segment between the long arms of chromosomes 4A and 5A, only 13 (4AS-4BS-4DS) and 35 (4AL-4BL-4DL) gene triplets were identified to be shared between the short and long arms of the homoeologous group 4 chromosomes, whereas 121 and 100 triplets were formed by genes from chromosome arms 4AL-5BL-5DL and 5AL-4BL-4DL, respectively. A

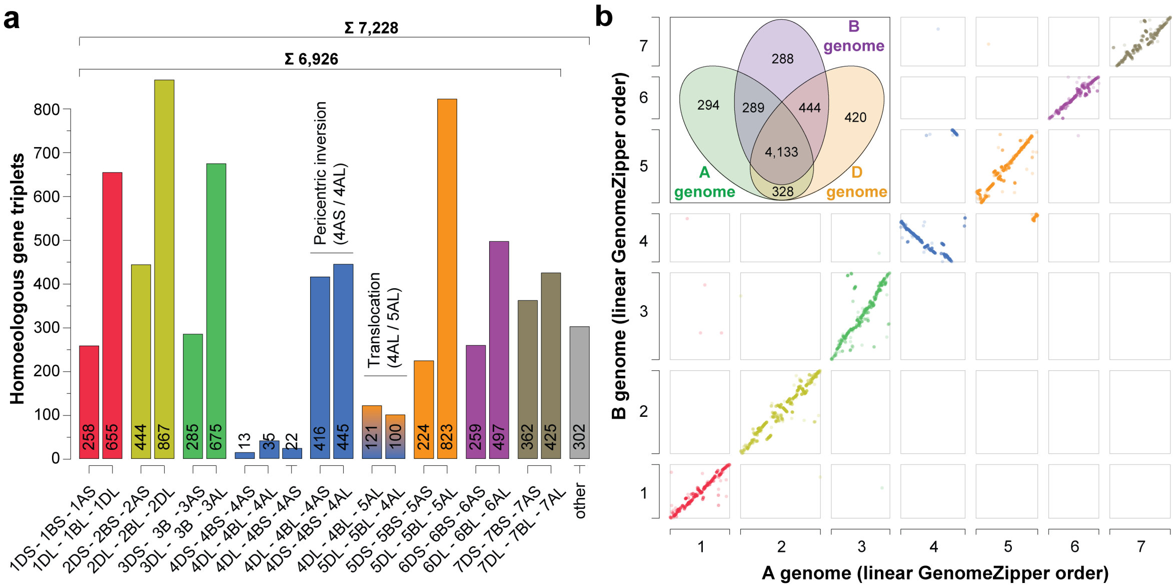


Fig. 4.12. Structural and functional characteristics of identified homoeologous gene triplets. **a,** Number of homoeologous gene triplets for each linkage group. **b,** Structural comparison of the distribution of homoeologous genes. The Venn diagram counts number of triplets anchored in and visualizes overlap in between the GenomeZippers for the A, B and D genomes. The dotplot depicts the linear ordering of homoeologous genes between the GenomeZippers of the A and B genomes. The corresponding structural comparisons between the A and D genomes and B and D genomes are shown in Fig. A.1.

[(4)] I gratefully acknowledge Sapna Sherma for implementation and identification of homoeologous gene triplets.

total of 416 and 445 gene triplets consisted of genes from chromosome arms 4AS-4BL-4DL and 4AL-4BS-4DS, respectively, which mirrored the two pericentric inversions happened between the short and long arms of chromosome 4A.

Furthermore, at least one gene was anchored in the wheat A, B and D wheat GenomeZippers for 6,196 (86%) triplets and all three homoeologs for 4,133 (57%) triplets (Fig. 4.12b). Most homoeologous genes were positioned in high co-linearity except for the previously described chromosomal re-arrangements involving chromosomes 4A, 5A and 7B *(63,314,315)* (Figs. 4.12b and A.1). However, small-scale interruptions in the micro-synteny were also evidentin pairwise comparisons of the ordering of homoeologs between chromosomes.

Distribution of protein function categories among homoeologous gene triplets

Besides genome-wide structural representativeness of the identified homoeologous triplets, I also evaluated the functional representativeness of protein function categories among single-copy homoeologous triplets. Therefore, I compared the general distribution of molecular function and biological process gene ontologies for proteins forming homoeologs gene triplets against the entire wheat gene catalogue utilizing GOSlim analysis *(316)*, which projects the granular gene ontology classification onto a more broad abstraction level [R/Bioconductor GSEAbase package (version 1.24.0) using the provided mapping file ”goslim_plant.obo”]. To test if the defined homoeologous triplets were biased towards specific GOSlim categories, I also performed a permutation test and compared the observed distributions against those computed for random selections of homoeologous triplets from the entire bread wheat gene space (1,000 iterations). Notably, the relative distributions of molecular function (Fig. 4.13) and biological process gene ontology categories (Fig. A.2) did not deviate for homoeologous triplets, the entire wheat gene space and the permu-

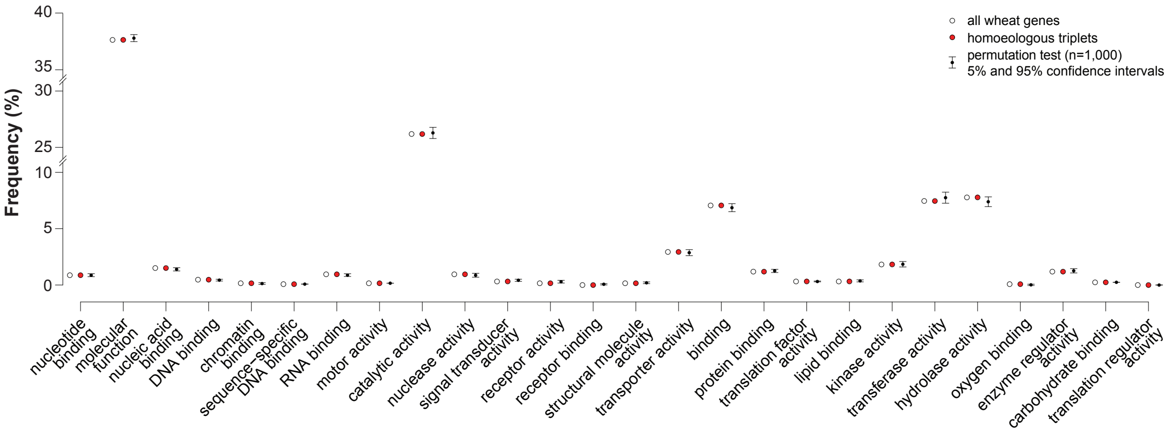


Fig. 4.13. Distribution of molecular function categories for homoeologous gene triplets and the entire wheat gene repertoire.

Comparison between the distributions of molecular function categories for homoeologous gene triplets, the entire wheat gene repertoire and a permutation of randomly selected gene triplets (1,000 iterations). Corresponding distributions for biological processes are shown in Fig. A.2.

tation test. This suggested that the defined set of homoeologous triplets constituted a robust and representative framework for genome-wide comparisons among wheat genomes.

4.3.4 Composition of wheat gene families

To test the extent of gene conservation across homoeologous chromosomes, the 133,090 predicted HC genes were clustered into protein families by sequence similarity using TribeMCL *(317)*. This identified a total of 10,684 TribeMCL groups and 5,606 singletons, i.e. wheat genes without sufficient sequence homology to others[5]. I merged both sets into a total of 16,290 gene family groups, which contained between 1 and 2,996 genes with a geometric mean of 3.1 genes per group and a median of 3 genes per group, respectively. Furthermore, I evaluated the genome composition of predicted wheat gene families and, therefore, converted the gene family grouping into a binary matrix. This matrix encoded the composition of the TribeMCL groups with respect to presence and absence of family members on individual chromosome arms. Then, I determined conservation in the gene family structures by hierarchical clustering analysis of the matrix with the `pvclust`-function *(318)* in R (binary distance and the "average" linkage method as well as 500 bootstrapping replications in order to estimate the uncertainty in the hierarchical clustering) (Fig. 4.14).

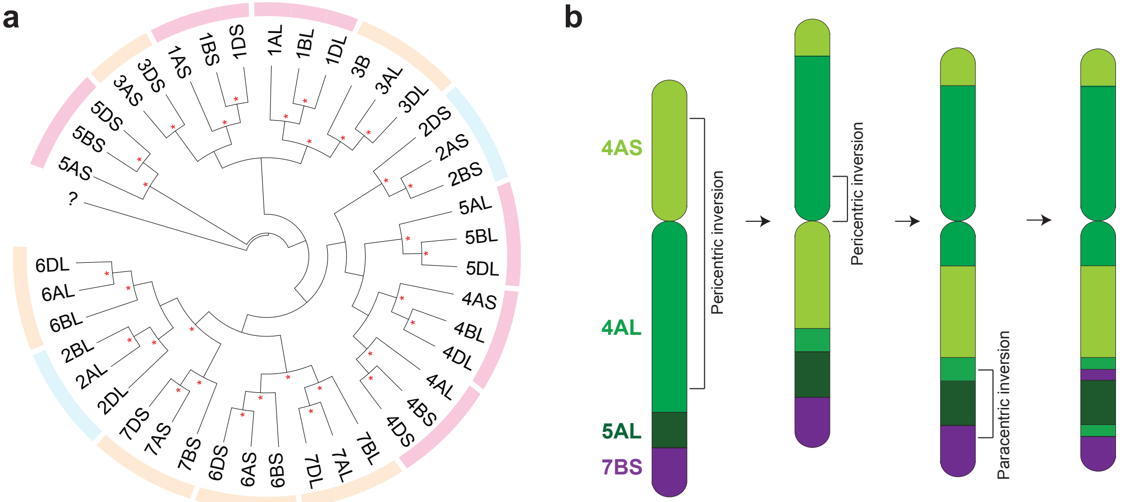


Fig. 4.14. Composition of wheat gene families.
a, Genome- and chromosome arm contribution to gene family clusters was subject to hierarchical cluster analysis. Color coding in the outer ring indicates relatedness of the respective branches. Red stars mark significant edges (boot strapping values >0.95). The "?" represents the set of wheat transcripts not found in the CSS assembly. **b,** Evolution and structure of chromosome 4A, which structure has been shaped through two translocation events (5AL to 4AL and 7BS to 4AL, respectively) and three subsequent peri- and paracentric inversions *(63,314,315)*. The coloring indicates chromosome-of-origin for individual chromosomal segments. [*The evolution of chromosome 4A is based on schematic drawings in (63,315).*]

[5] I gratefully acknowledge Manuel Spannagl for computing the TribeMCL clustering and providing the gene family groups for further analysis.

With the exception of chromosome 4AL, all chromosome arms clustered with their corresponding homoeologous counterparts (Fig. 4.14a). However, the pattern of clustering observed for homoeologous chromosome group 4 was consistent with the patterns observed for the homoeologous gene triplets (Fig. 4.12). It reflected a known pericentromeric inversion, which interchange a segments native to the short and long arms, respectively, and two translocations of segments from chromosome arms 5AL and 7BS *(63,314,315)* (Fig. 4.14b).

Considering only the grouping of homoeologous chromosome arms, all possible cluster topologies between genes in the A, B and D genomes were apparent. For example, 7A and 7D shared more homoeologues than they shared with 7B. In contrast 5B and 5D shared more genes than they did with 5A or 2A and 2B shared more genes than they did with 2D. Thereby, notably, the topologies occurrence in unbalanced frequency. Whereas five and six homoeologous groups formed the topologies *A(B,D)* and *B(A,D)*, respectively, only for the short and long arms of the homoeologous chromosome group two the A genome and B genome showed highest conservation in gene content. Overall, these patterns indicated that A and B chromosomes were most different with respect to gene content, with the D chromosomes being about equally similar to A as to the B chromosomes. This finding was consistent with other phylogenetic studies by Marcussen *et al. (69)* on basis of the genomic resources generated in this thesis.

4.4 Alternative splicing in bread wheat

Alternative splicing (AS) of precursor (pre-)mRNAs constitutes a major transcriptional mechanism, which is common to all eukaryotic organisms, to increase the functional diversity of the proteome *(308,319)*. For example, the generation of multiple splice variants from one gene has shown to be important in the response to environmental stresses allowing efficient and rapid adaption to changing conditions *(184,307)*. Furthermore, transcriptome-wide studies have not only shown that AS provides myriad of additional protein variants, emerging evidences suggest that AS plays a major role in post-transcriptional gene regulation and impacts transcript stability, translation and transcript localization through, for example, generating different ezymatic products *(320)*, microRNA-mediated gene regulation *(25,321)* and nonsense-mediated decay *(184,307,322,323)*. Previously, genome-wide analyses of AS have been hampered by the lack of comprehensive, sufficient deep and multi-tissue transcriptome data sets. This has dramatically been changed with the emergence of high-throughput mRNA-seq technology. Recently, more than 60% of multi-exon genes have been reported to be alternatively spliced in *Arabidopsis* under normal growth conditions *(184)*. However, this might be a conservative estimate as AS is often regulated specifically in individual tissues or in changing environmental conditions *(324)*.

Usage of RNA-seq transcriptome data of five organs (leaf, grain, root, stem and spike) revealed high abundance of AS in the bread wheat genome (Table 4.7). A total of 61,168 alternatively spliced HC wheat genes were predicted and a total of 262,114 distinct splicing variants structurally annotated. In the following, the observed splicing patterns will be investigated, AS compared among homoeologous wheat genomes and the conservation and impact of potential post-transcriptional regulatory mechanisms discussed.

4.4.1 Distribution of alternative splicing in bread wheat

As large variations in the structural characteristics of genes from different confidence classes were evident (Section 4.3.1). Therefore, the following analysis were restricted to the highest confidence class level HC1 including protein-coding, full-length genes (HC1), in order to avoid biased observations due to fragmentation and incomplete assembly of transcript structures.

Genome-wide distribution of alternative splicing

Overall, a comparable degree of AS was detected for the A genome (69% alternatively spliced genes), B genome (68%) and D genome (69%) (Fig. 4.15a). Also, a similar number of transcript variants per gene was predicted across genomes (A genome: 3.5, B genome: 3.5 and D genome: 3.6) (Fig. 4.15b). However, slight variations were evident in proportion of alternatively spliced genes [63% (3DS) to 75% (5AL)] and in the mean number of splicing variants per locus [3.0 (1AS) to 4.0 (7DL)] for individual chromosome arms.

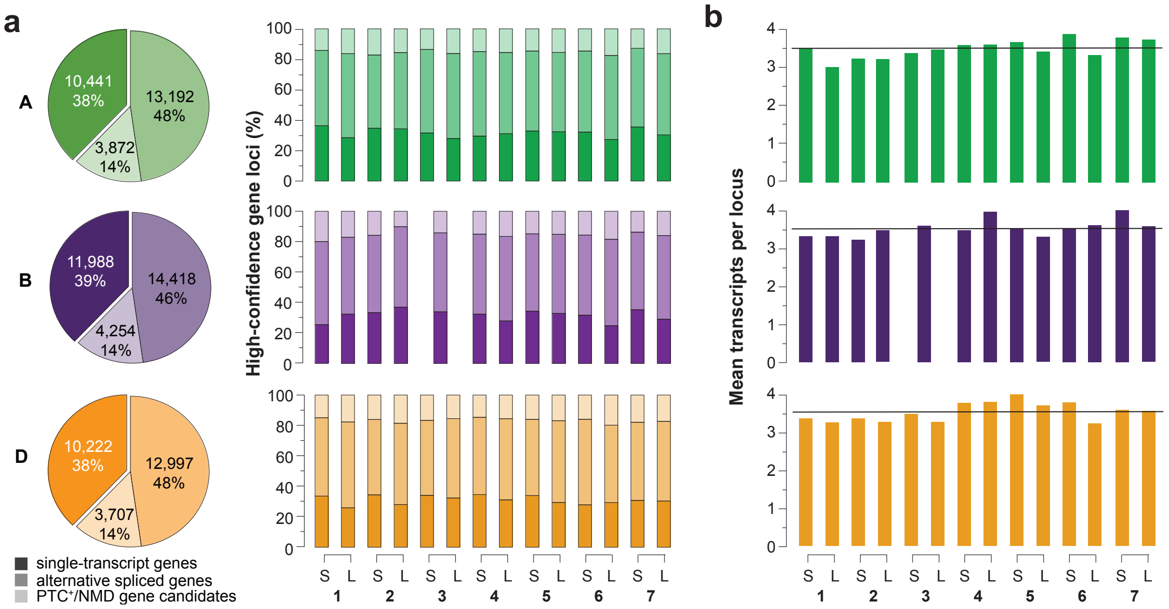


Fig. 4.15. Distribution of alternative splicing across genomes and chromosome arms.
For the A, B and D genome and individual chromosome arms the figure visualizes **a,** the number of HC1 wheat genes with single transcript, alternative transcripts or with at least one PTC$^+$/NMD candidate and **b,** the mean number of annotated transcripts per gene for individual chromosome arms. Black lines depict the mean over all genes of the respective genome.

Distribution of splice types

The different types of AS events were analysed for the HC1 wheat genes with the ASTALAVISTA software package *(325)*. Intron retention was found to be the most common type explaining approximately one quarter of all splicing events (Fig. 4.16). The next most frequent splicing events were alternative 3' acceptor sites (19%) and 5' donor sites (16%). Exon skipping was only rarely observed in wheat (6%). Additionally, this analysis revealed a large number of complex constructs built up by the combination of different single splicing events like, for example, multiple skipped exons. However, the observed splicing events were similar frequent for the wheat A, B and D genomes and, moreover, largely consistent with observations in *Arabidopsis (184,307,326)*, which indicated high conservation of exon splice types across the angiosperms.

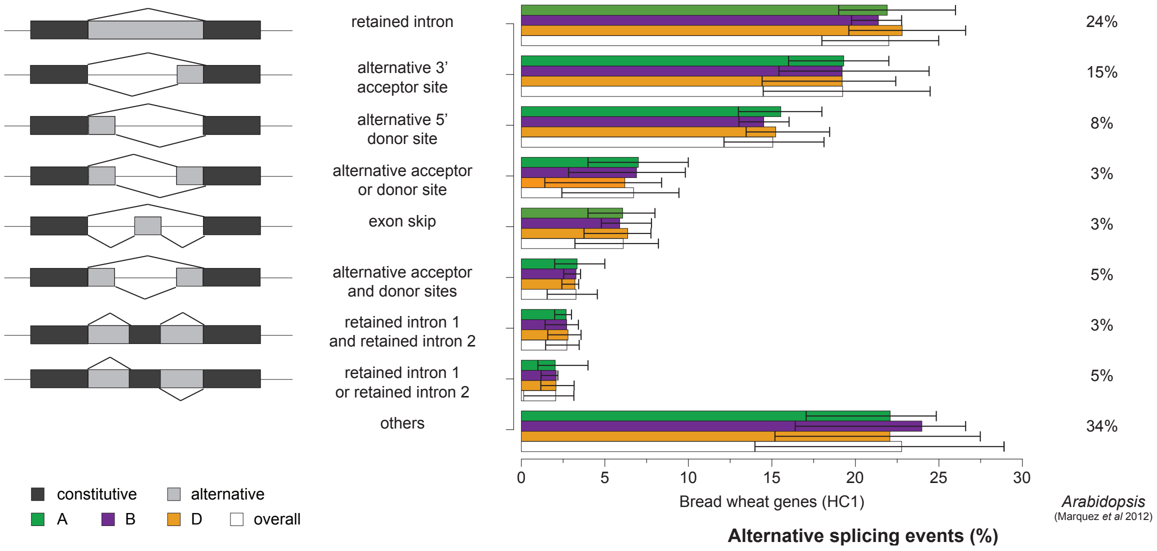


Fig. 4.16. Frequency of alternative splicing events in bread wheat.
Frequency distribution of the most frequent types of alternative splicing events in the predicted wheat transcripts across the A, B and D genomes. For comparison the observed frequencies for the *Arabidopsis* gene annotation are shown *(184)*.

Conservation of alternative splicing across homoeologous wheat genes

So far, no global differences in the fraction of alternatively spliced genes were evident. Therefore, I analysed if strictly single-copy homoeologous gene triplets were affected differentially by AS. Only considering the 3,797 homoeologous triplets that were formed by HC1 wheat genes (Section 4.3.3), all three homoeologs were alternatively spliced in 2,829 cases (76%) (Fig. 4.17). Only 14% and 7% of the analysed triples showed a mixture of normally and alternatively splicing genes in the A, B and D genomes, respectively. Significant differences were observed compared to a permutation test using randomly defined gene triplets (1,000 iterations). All three homoeologous genes to be alternatively spliced was observed for approximately one third of the randomly generated triplets, which assumed complete independence among genes forming triplets. On the contrary, significantly more genes with non-balanced alternative splicing patterns among homoe-

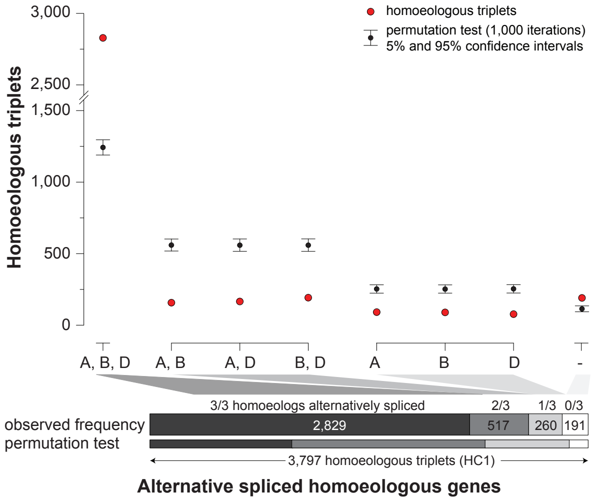


Fig. 4.17. Conservation of alternative splicing among homoeologous gene triplets.
Number of homoeologous triplets for which multiple alternative transcripts variance were predicted for all three homoeologs, for two or one homoeologs and any homoeolog. Only homoeologous gene triplets formed by HC1 wheat genes were considered. The observed frequencies were compared to a the frequencies observed for randomly formed triplets (1,000 permutations).

ologs than observed for bread wheat would be expected by assuming a random occurrence of AS.

4.4.2 Analysis of post-transcriptional gene expression regulation

Regulation of eukaryotic gene expression is a complex network of myriad different mechanisms and pathways including transcription, RNA processing and export, translation as well as protein folding *(25,320–323)*. Strict control of the involved participants and individual steps is particularly crucial for an organism's vitality and for orchestrating and maintaining all cellular processes like, for example, adaptation to changing environmental conditions or response to external stimuli *(327,328)*. The nonsense-mediated decay (NMD) pathway is one of these important quality-control mechanisms and detects, targets and degrades alternatively spliced transcripts, which contain premature termination codons (PTCs) *(323,329)* (Fig. 4.18). Those PTC^{+} transcripts arise by a nonsense stop codon that occurs before the authentic stop codon of the functional transcript and, consequently, encode truncated proteins. The NMD surveillance pathway ensures removal of potentially non-functional transcripts. Furthermore, the generation of premature stop-codon containing mRNAs has been also demonstrated to be an important post-transcriptional regulator of gene expression, especially in response to environmental stresses *(307,330,331)*. In contrast to transcriptional regulation of gene expression, which controls the transcription of genes into pre-mRNA, this type of gene expression regulation is controlled by the splice environment and has been termed "regulated unproductive splicing and translation" (RUST) *(332)*. Thus, the generation of alternatively spliced transcripts, which are differentially subjected to NMD,

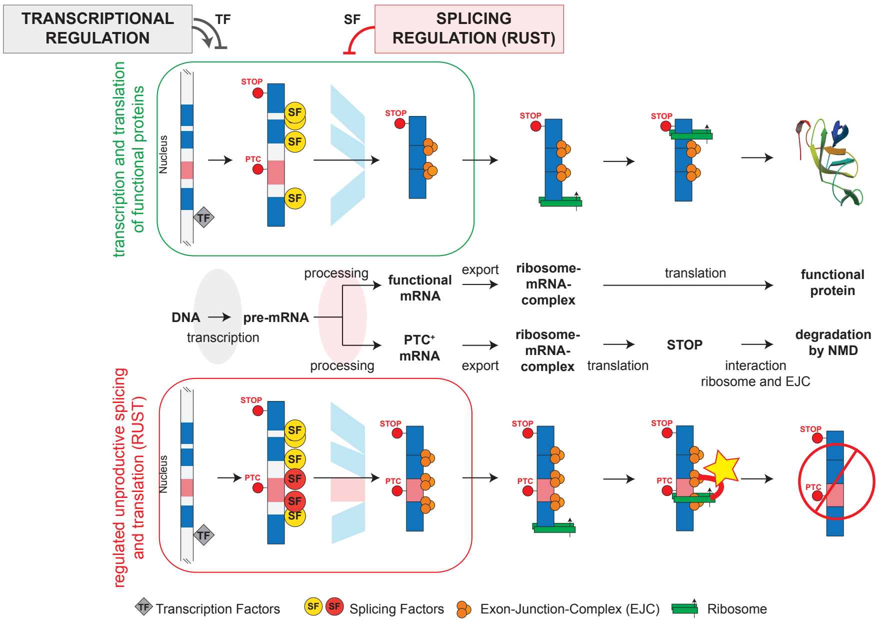

TRANSCRIPTIONAL REGULATION
TF
SF
SPLICING REGULATION (RUST)
transcription and translation of functional proteins
Nucleus
STOP
SF
PTC
TF
processing
functional mRNA
export
ribosome-mRNA-complex
translation
functional protein
DNA
pre-mRNA
transcription
PTC+ mRNA
STOP
degradation by NMD
interaction ribosome and EJC
regulated unproductive splicing and translation (RUST)
Transcription Factors
Splicing Factors
Exon-Junction-Complex (EJC)
Ribosome

Fig. 4.18. Gene expression regulation by unproductive splicing and translation. Simplified scheme of gene expression regulation via unproductive splicing and translation (RUST). On contrary to transcriptional regulation, which controls the transcription of DNA into pre-mRNA via activation or repression by transcription factors, splicing factors determine exon usage and, thus, the generation of productive (upper pathway) mRNAs or non-productive (lower pathway) mRNAs, respectively. Exon-junction complexes were placed to the splice sites during pre-mRNA processing and mark gene structure. Whilst productive mRNAs are translated into functional proteins, in the RUST pathway the ribosome stops at the (nonsense) premature termination codon (PTC). Release factors interact with the remaining exon-junction complexes and trigger degradation of the PTC$^+$ transcript by the nonsense mediated decay (NMD) pathway. [*This figure is based on background information of (332), which has been provided by the authors on http://compbio.berkeley.edu/people/ed/rust.*]

represents an additional layer of complexity regulating protein expression.

Identification of PTC$^+$ transcripts

The molecular mechanisms of NMD have been detailed described *(323,329,333)* and PTC$^+$ transcripts were defined by occurrence of a stop codon more than 50 nucleotides upstream of the following three-prime exon/exon splice junction *(334–336)*. By using this classification criterion, I screened for potential PTC$^+$ transcripts on basis of the exon structures and ORF information of the 261,881 transcripts predicted for alternatively spliced HC wheat genes.

In total, 37,196 transcripts (14%) contained a PTC and might potentially be degregated by NMD. As recent studies in *Arabidopsis* suggested that transcripts with retained introns are not sensitive to NMD *(307)*, I filtered out a total of 9,330 transcripts, of which the premature stop was caused by intron retention. This resulted in the final computational prediction of 27,866 PTC$^+$ transcript candidates (11%), which were annotated for 14,972 HC gene loci (23%) (Table 4.8). Considering only full-length wheat gene predictions (HC1), comparable levels of PTC$^+$/NMD sensitive gene loci were detected among all wheat genomes (Fig. 4.15a). A total of 3,872 out of

17,064 alternatively spliced genes of the A genome were classified potentially to be regulated by PTC$^+$/NMD (23%), 4,254 out of 18,402 genes of the B genome (23%) and 3,707 out of 16,704 genes of the D genome (22%). On the level of individual chromosome arms between 19 and 24% were classified as potential PTC$^+$ transcripts on level of individual chromosome arms.

Table 4.8. Alternative splicing and transcripts containing PTCs across high-confidence gene loci.

	HC1	HC2	HC3	HC4	Σ
General statistics of alternative splicing in high-confidence supported gene loci					
Predicted HC genes	55,429	14,367	15,475	39,110	124,201
Predicted transcripts at high-confidence genes	194,624	37,116	31,957	61,450	325,147
Genes with alternative transcripts	38,059	7,916	6,465	8,728	61,168
Predicted transcripts derived from genes with alternative splicing	177,434	30,665	22,947	31,068	262,114
Premature stop codon analysis					
Predicted transcripts used for PTC analysis[a]	177,338	30,630	22,919	30,994	261,881
Transcripts without PTC	153,436 (87%)	26,370 (86%)	19,703 (86%)	25,176 (81%)	224,685 (86%)
Transcripts containing PTC	23,902 (13%)	4,260 (14%)	3,216 (14%)	5,818 (18%)	37,196 (15%)
PTC (intron retention)	6,168 (3%)	1,071 (4%)	749 (3%)	1,342 (4%)	9,330 (4%)
PTC$^+$ transcript candidates	17,734 (10%)	3,189 (10%)	2,467 (11%)	4,476 (14%)	27,866 (11%)
Genes with PTC$^+$ transcripts	8,876 (23%)	1,649 (21%)	1,308 (20%)	2,139 (25%)	13,972 (23%)

[a] Only transcripts were used for which a protein sequence was predicted.

Conservation of PTC$^+$ transcripts for homoeologous wheat genes

Furthermore, the presence and the conservation of PTC$^+$ transcripts were elucidated among genes forming homoeologous triplets (Section 4.3.3). While no evidence of PTC$^+$/NMD was found for 2,294 triplets (60%), at least one homoeolog of 1,503 triplets encoded a PTC$^+$ transcript (40%) (Fig. 4.19). Only one PTC$^+$ transcript was detected for the majority of these triplets [833 triplets (55%)]. However, for 267 of these triplets (18%) all three homoeologous genes encoded PTC$^+$ transcript(s), thus showed evidence for post-transcriptional regulation by the RUST pathway across genomes. This finding was significant different to the distribution observed by a permutation test, in which complete independence of PTC$^+$/NMD sensitivity was assumed for homoeologous genes forming triplets (1,000 iterations and P value <0.05). Compared to an observed presence of a PTC$^+$ transcript for all three homoeologes genes of a triplet in 17% of the cases, less than 1% would be expected by chance. On the contrary, this test revealed significantly more homoeologous gene triplets without PTC$^+$ transcripts for all three homoeologs and less triplets with only a genome-specific encoded PTC$^+$/NMD gene candidate.

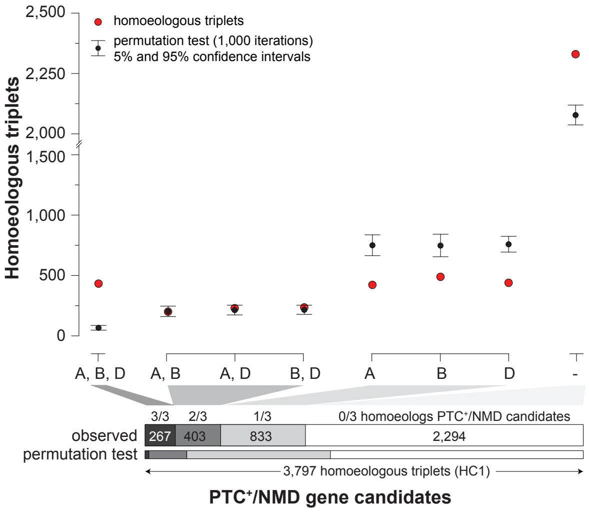


Fig. 4.19. Conservation of PTC$^+$ /NMD gene candidates among homoeologous triplets. Number of homoeologous triplets for which all three, two, one or any homoeologs were classified as PTC$^+$/NMD genes. Only homoeologous gene triplets formed by HC1 wheat genes were considered. The observed frequencies were compared to a the expected frequencies for complete independence between genes of a triplet as tested by randomly formed triplets (1,000 permutations).

4.5 Discussion

Accompanied by on-going improvement of next generation sequencing technology, which allowed obtaining comprehensive genomic sequence resources with decreasing costs, chromosome (arm) sorting has been a milestone in wheat genomics and allowed the IWGSC to construct a draft genome sequence assembly for hexaploid bread wheat. By using reference protein information of closely related grass species and a comprehensive wheat RNA-seq data set, I developed an extrinsic gene annotation pipeline and structurally annotated the bread wheat genome. Thereby, I took advantages of the physical separation of homoeologous genomes into single chromosome arm bins and investigated genome-wide structural relationships among the A, B and D genomes.

4.5.1 A comprehensive annotation of protein-coding bread wheat genes based on extrinsic sequence information

The implemented gene finding and annotation pipeline enabled predicting a genome-wide set of bread wheat genes. Using extrinsic sequence information 133,090 high-confidence protein-coding wheat genes were identified, of which 124,201 genes (93%) were structurally defined the chromosome survey sequences and assigned to individual chromosome arms. The remaining 7% corresponded to wheat transcript sequences not represented in the CSS assembly (Table

4.6). During annotation I conducted stringent homology-based confidence analysis to consider technical fragmentation of genes onto multiple contigs in the CSS assembly as well as to identify and distinguish between full-length protein-coding genes and pseudogenes or gene fragments. Thereby, I further subdivided the HC genes into four confidence levels based on sequence coverage to orthologous proteins. Overall, 55,249 of the predicted wheat genes located in the CSS assembly (44%) were assigned to the highest confidence class (HC1) and spanned at least 70% of the length of the supporting evidence. A total of 29,842 and 39,110 wheat genes were further identified with medium coverage of orthologous genes [<70% and $\geq$30% coverage (HC2 and HC3)] and with very low coverage [<30% coverage (HC4)], respectively. Moreover, a homology-based approach resulted in the definition of more than 7,000 homoeologous gene triplets, which provided a suitable framework for in-detail analysis of homoeologous relationships between the A, B and D wheat genomes.

Based on the number of identified HC1 genes and sequence coverage of high-quality wheat fl-cDNA sequences by different independent genomic resources, the bread wheat genome was estimated to contain approximately 106,000 genes (Fig. 4.9). This estimate corroborated previous findings of this thesis on basis of WGS sequencing and the orthologous group assembly (Section 3.2.3) and was consistent with estimates of other studies ranging between 32,000 and 38,000 for diploid wheat genome *(43,44,63,241)*.

The predicted gene set represented almost the entire bread wheat genome. Ninety-six percent of publicly available wheat ESTs were detected, with 89% of the ESTs already by HC genes only (Fig. 4.8). Additional, independent confirmation of gene structure prediction was also emerging from proteomics analyses of wheat proteins *(70)*. From 63 genes tested, 50 (81%) were confirmed, eight (13%) provided evidence for alternate structures and five (8%) were absent in the structural gene calls[6].

4.5.2 Identification of thousands of gene fragments, pseudogenes and non-coding transcriptional active regions in the wheat genome

Although the fragmentation of genes into two or more contigs in the CSS assembly, which could not be detected by the implemented template-based joining algorithm, has to be considered, abundance of gene fragments and pseudogenes in the wheat genome *(199,225)* impeded gene prediction and most probably cause an overestimated number of genes in the entire HC gene set. A proportion of low (HC4) and medium (HC2 and HC3) confidence genes might have represented true but incompletely defined genes. However, declining sequence conservation to orthologous proteins (Fig. 4.6) and decreased representation in the public wheat HarvEST database *(203)* indicated that with decreasing protein-coding-confidence, the high-confidence gene sets HC2 to HC4 included also a substantial number of deteriorated gene fragments and pseudogenes

[6] I gratefully acknowledge Matthew Bellgard, Brett Chapman and Rudi Appels (Murdoch University, Australia) for the proteomic analysis and evaluation of the predicted protein sequences.

(Fig. 4.8). Especially, the HC4 gene set most likely accumulated relatively young pseudogenes, whose protein-coding sequence have not been sufficient degenerated to be classified into the LCS gene set. Median alignment similarity to the respective reference protein was reduced for HC4 genes, of which the majority had 10% to 20% reference gene coverage. This was consistent with previous observations discussed in chapter (Section 3.3) of this thesis, showing that repeat-associated wheat sub-assemblies formed ”stacks” after the orthologous group assembly.

Furthermore, 95,398 LCS gene loci were identified with homology to plant reference species, but at significant reduced protein conservation levels. LCS gene loci were less frequently located in syntenic conserved regions (Fig. 4.11). The locus sizes of LCS genes were substantially shorter (519 bp) compared to high-confidence genes (1,112 bp) (Table 4.7). More than two third of the LCS genes were single exon genes and on average 1.28 alternative transcripts structures were annotated per locus revealing that low-confidence genes were less affected by alternative splicing compared to bona-fide protein-coding genes (30% single exon genes and 2.62 alternative transcripts per locus). On the contrary, almost doubled median exons size for LCS genes (264 bp for LCS genes compared to 168 bp for HC genes) suggested that these loci might represent non-processed pseudogenes and originated from retro-transposition of a RNA intermediate back into the genome *(270)*. Taken together, although a proportion of HC4 and LCS genes might represent fractions of functional gene, these observations indicated that a majority of these gene sets most likely represented non-functional genes, gene fragments or deteriorated (pseudo-)genes, which resulted from generation and amplification by DNA transposons, retroelements or double-strand break repair *(199,224,225,275)*.

A total of 728,459 predicted loci did not share any significant homology to plant proteins (USL confidence class) or completely lacked a reasonable open reading frame (NCL confidence class) (Table 4.7). In part, these predictions resulted from the repetitive nature of the bread wheat genome. Ultra-short seeds in the CSS assembly caused spurious alignments of reference proteins or wheat cDNAs sequences, which bridged repeat-masked genomic sequences. The translated peptide sequences of these structures had large proportion of repeat-masked sequences and resulted in amino acid sequences without functional relevance. However, a substantial proportion of the USL and NCL loci showed transcriptional evidence. While these gene set might also included potential species-specific genes, they more likely represented novel (non-protein coding) transcriptional active regions, which have been also described for numerous species including both, plants *(191,309)* and animals *(337)*.

4.5.3 Dynamics of the bread wheat genome

Overall, the gene repertoires in the A genome (40,253 HC genes) and D genome (39,425 HC genes) were of similar size, both exceeded by the gene catalogue of the B genome (44,523 HC genes). In contrast, considerable differences were apparent among individual homoeologous chromosomes and chromosome arms with variations in gene counts (Fig. 4.10) and in gene density (Fig. 4.11). Fivty-three percent of the genes were positioned into syntenic conserved regions and were anchored to the syntenic framework of genes from *Brachypodium*, rice and sorghum in the wheat GenomeZipper, which was consistent with conservation of synteny in other grasses *(61,62)*. Notably, large differences in the syntenic conservation among individual chromosome arms were evident. A generally higher conservation was found in the D genome (58%) compared to the A and B genomes (approximately 50%). Although differences in the underlying marker map that were used for construction of the individual GenomeZipper of each genome had to be considered, the observed differences might already be set in the diploid progenitors of the wheat A-, B- and D-genomes and, thus were inherited to tetra- and hexaploid wheat. Alternatively, these findings reflected the evolutionary history of hexaploid wheat and may indicating an increased disruption of synteny for the A- and B-genome chromosomes during common polyploid evolution. Moreover, the different conservation rates for chromosomes indicated that the control of genome composition act locally on distinct chromosomes, chromosome arms or segments within chromosomes rather then on the level of entire homoeologous genomes.

Protein sequence-based clustering was used to group the predicted HC genes into gene families. The comparison between the expected gene family sizes of diploid grass genomes (in average 1.4 genes per family) with the observed sizes of the gene families in bread wheat (in average 3.1 genes per family) allowed estimating the hexaploid-to-diploid gene retention rate to approximately 2.2. This finding largely corroborated the previous estimates on the basis of wheat WGS sequencing and the orthologous-group assembly (Section 3.2).

While analysis of the paleopolyploid maize genome *(111)* has been shown preferential loss of genes from one genome, the observed patterns for bread wheat did not indicate favoured genome dynamics acting on a particular wheat genome. Assuming almost similar genome sizes of the diploid progenitor genomes of bread wheat *(43,44)*, the generally balanced gene content across wheat genomes suggested structural autonomy as a result of prevented inter-genome recombination due to restrained pairing of homoeologous chromosomes during meiosis (Section 1.3.1). However, incongruence in gene family composition among homoelogous chromosome arms indicated a non-uniform interchange between wheat genomes (Fig. 4.14). Rather then linear evolution of the A-, B- and D-lineages in the Triticeae, this could be explained by bifurcating evolutionary relationships among the diploid genome donors as shown recently in phylogenetic analysis of Marcussen *et al. (69)*. Additional evidences have also suggested that Triticeae genomes show a dynamic genome composition including non-linear, reticulated evolution and, at least partially, have been shaped by large-scale introgressive events or incomplete lineage sorting *(69,71,73)*.

4.5.4 A highly complex and conserved alternative splicing landscape

Usage of RNA-seq technology allowed investigating the alternative splicing landscape of bread wheat and the structural definition of 325,147 distinct transcript variants. Fourty-nine percent of HC genes were found to be alternatively spliced with on average 2.62 transcripts per locus (Table 4.7). These observations were largely consistent with observations in barley *(191)* and *Arabidopsis (184,307)* and confirmed that AS is a major regulatory mechanism, which increases transcriptome and proteome complexity and diversity. Almost 70% of the genes of the most complete gene class (HC1) were alternatively spliced with on average 3.5 transcripts per locus. This allowed extrapolating the wheat transcriptome to contain more than 300,000 protein-coding transcripts. None of the homoeolog genomes was predominately affected by AS (Fig. 4.15) or showed differential usage of splicing events across the homoeologous wheat genomes (Fig. 4.16). Intron retention was the most common AS event (24%), followed by alternative 3' (15%) and 5' donor sites (5%), which was consistent with studies in *A. thaliana (184,307,326)* and indicated highly conserved splicing patterns within the plant kingdom and over 150 mio years of evolution.

The NMD-pathway is an important surveillance mechanism, which rapidly detects and degrades aberrant RNA transcripts like, for example, PTC^{+} transcripts that encode for truncated proteins *(332)*. PTC^{+}/NMD also constitutes an important post-regulatory transcriptional mechanism, which has been shown to act often in environmental stress response *(307,330)*. A total of 27,866 AS-transcripts (9%) contained premature termination codons and were located at 13,972 high-confidence genes loci (11%) (Table 4.8), which was largely comparable to studies in plants *(191,323)* and animals *(338)*. Across genomes, similar degree of AS and PTC^{+}/NMD sensitivity was observed (Fig. 4.15), which was also conserved among homoeologous single-copy gene copies (Figs. 4.17 and 4.19). This suggested that both mechanisms were maintained in the hexaploid genome and have already been evolved in the diploid progenitor genomes before hybridization or, probable, before specification of individual genome lineages. These findings contradicted with the "spurious transcript" model, which hypothesizes the NMD pathway is exclusively a quality control mechanisms removing nonsense transcripts that are costly-to-make, thus are potentially deleterious for the cells fitness *(339)*. Conservation of PTC^{+}/NMD more supported the "regulatory transcript model" that concede post-transcriptional regulatory functions to the PTC^{+}/NMD machinery, which modulates gene expression via splicing factors *(332,339)*. This finding corroborated recent observations in both mammals *(340)* and plants *(307)*.

4.6 Conclusions

Working on the IWGSC chromosome sequence survey assembly facilitated to identify nucleotide and protein sequences and transcript structures for more than 90% of bread wheat genes. The generated genomic resources have been made publicly available for visualization and download in the EnsemblPlants web portal (http://plants.ensembl.org/Triticum_aestivum) hosted by the EMBL-EBI. This thesis revealed a highly complex genome structure, which was characterized by high abundance of low-confidence genes including non-coding transcribed regions as well as deteriorated gene fragments and pseudogenes in addition to high-confidence protein-coding gene loci. The determined protein sequence set allowed elucidating gene family sizes and compositions on a chromosome arm level in bread wheat, which supported reticulate evolution in the Triticeae. Gene families were reduced in bread wheat confirming previous findings using whole genome shotgun sequencing and the outcome of the orthologous group assembly. No bias towards predominant retention or loss of genes for one of the three homoeologous genomes was observed, suggesting a high level of plasticity of the hexaploid wheat genome, while, simultaneously, each homoeologous wheat genome is autonomously maintained. However, differences in syntenic conservation and gene density on a chromosome arm level indicated molecular mechanisms to shape differentially homoeologous chromosome arms or individual chromosomal regions. Moreover, this work highlighted alternative splicing to be an additional layer of complexity, which largely increase the diversity of the bread wheat transcriptome. Conservation of splicing patterns and potential premature termination codon-containing transcripts across genomes supported the ”regulatory transcript model” attributing regulatory functions to the splicing machinery, which have emerged before polyploidization and are common for the *Triticum* genome lineages.

The generated data resources provide a suitable genomic framework for myriad analysis aiming at understanding the key mechanisms that shape the genome structure of allohexaploid bread wheat. Together with the putative chromosomal ordering, the predicted gene catalogue is of high value for targeted breeding to identify the genetic elements for the improvement of agronomic and industrial important traits of one of the most important crops worldwide.

Chapter 5

The transcriptome of hexaploid wheat during endosperm development

The previous chapters of thesis focussed on investigating the impact of polyploidization on genome content and structure of bread wheat revealing pronounced retention and structural conservation across homoeologous genomes. However, with bringing together multiple genome sets, polyploidization is one of the most challenging events in an organism's evolution (Section 1.3). Such a "genome shock" *(46)* has been shown to result in alterations of the regulatory mechanisms orchestrating inter- and intra-genome gene expression, balancing regulatory elements and accurately controlling protein levels for a highly redundant gene set *(101,107,114,121,126)*. Furthermore, analysis of synthetic polyploids and paleopolyploids have demonstrated that both, genetic *(120)* and epigenetic modifications *(118,120,124)*, might result in genome asymmetry and favoured expression of genes from a single genome *(115,116)*. in wheat, however, those studies were based on a limited number of genes *(121,123,124)* and the extent and characteristics of gene expression divergence between genomes in different tissues have been largely unknown at the whole genome level.

Because of the agricultural and industrial importance of bread wheat (Section 1.2) researchers have put special interest in enhancing specific grain quality attributes and in the investigation of the genetic control of grain components. For allohexaploid wheat, partial or complete genome dominance has been found to affect various morphological and agronomic traits including grain protein content *(136)* and grain hardness *(341)*. However, a major impediment to a genome-wide understanding of transcriptional relationships among the homoeologous wheat genomes was the absence of a suitable reference genome sequence that enables measuring A, B and D genome-specific transcription. This restricted studies only to single genes *(342–344)* or onto a global analysis without homoeologous resolution *(345,346)*.

In a collaboration with a research team led by Prof. Dr. Odd-Arne Olsen of the Norwegian University of Life Sciences (Ås, Norway), this study made use of the genomic resources established by the IWGSC (Chapter 4) to investigate gene expression for three developmental stages in different cell types of wheat endosperm. Besides providing technical guidelines for the application of high-throughput RNA-sequencing to comprehensively analyse the transcriptome of one of the most complex plant genomes, patterns of spatiotemporal gene expression will be examined and functionally characterized on several levels in the following sections. Starting from a global prospective on the wheat endosperm transcriptome, specific functional aspects of grain development will be highlighted, potential key regulators and marker genes defined and co-expressed genes grouped into clusters with distinct expression profiles. In particular, this chapter will focus on investigating homoeologous-specific gene expression to gain novel insights into genome asymmetry and to elucidate positional effects on gene transcriptional regulation in a polyploid genome. Finally, a genome-wide catalogue of industrially important genes that are known affect wheat baking quality will be established in a targeted gene family analysis.

All methods and results shown in this chapter are part of following publications:

- **A chromosome-based draft sequence of the hexaploid wheat genome**
  The International Wheat Genome Sequencing Consortium (IWGSC)
  Science. 345(6194):1251788, 2014.
- **Genome interplay in the grain transcriptome of hexaploid bread wheat**
  M. Pfeifer[‡], K. G. Kugler[‡], S. R. Sandve, B. Zhan, H. Rudi, T. R. Hvidsten, IWGSC, K. F. X. Mayer and O.-A. Olsen
  Science. 345(6194):1250091, 2014.
  [‡] joint first authors

5.1 Developmental stages and major cell types of the nuclear endosperm

Starch constitutes about 65% to 75% of dry weight of mature cereal seeds *(347)*. Therefore, wheat grains belong to the major crop materials providing raw material for various industrial processes and contributing essentially to livestock feeding and human nutrition. Changing environmental conditions and worldwide population growth require an in-detail understanding of crop physiology and grain development to satisfy global demands and ensure food security *(15,19)* (Section 1.1). Cereal endosperm development partitions into three, partly overlapping phases: early development, differentiation and maturation *(217,348)*. During the first phase, early development, the endosperm origins from an initial triploid nucleus as a result from a double fertilization between a sperm cell nucleus and of two polar nuclei in the central cell of the embryo sac. Rapidly, the initially triploid nucleus starts to divide and proliferate without formation of cell walls, which leads to a multinucleate cell, the endosperm coenocyte (Fig. 5.1a). Subsequently, formation of a radial microtubule system and aveolation initiate cellularization of the coenocytic endosperm until completion of the central vacuole with cells, which is mostly completed approximately 3 to 6 days post anthesis (DPA) (Fig. 5.1b). In the next two phases, endosperm differentiation and maturation, the industrially important characteristics of the wheat grain are developed. Initial endosperm cells specialize into different cell types (Fig. 5.1c), expand, increase water content and accumulate starch and storage proteins (Fig. 5.1d). In the early development of endosperm cell type specification, which has been suggested to be mainly controlled via positional signalling, and endosperm cellularization overlap largely *(217)*.

The mature endosperm consists of four major cell types: transfer (TC) cells, aleurone (AL) cells, starchy endosperm (SE) cells and embryo-surrounding (ESR) cells, respectively (Fig. 5.1c). The ESR is located in the cavity of the developing endosperm in direct proximity to the embryo. Probably the ESR is involved in embryo nutrition and constitutes a physical barrier and communication zone between the embryo and the starchy endosperm, but the particular function of the ESR is unknown *(217)*. The ESR develops at an early stage and, corresponding to embryo-growth, shrinks subsequently at later stages *(350)*. Transfer cells are located in the basal of the

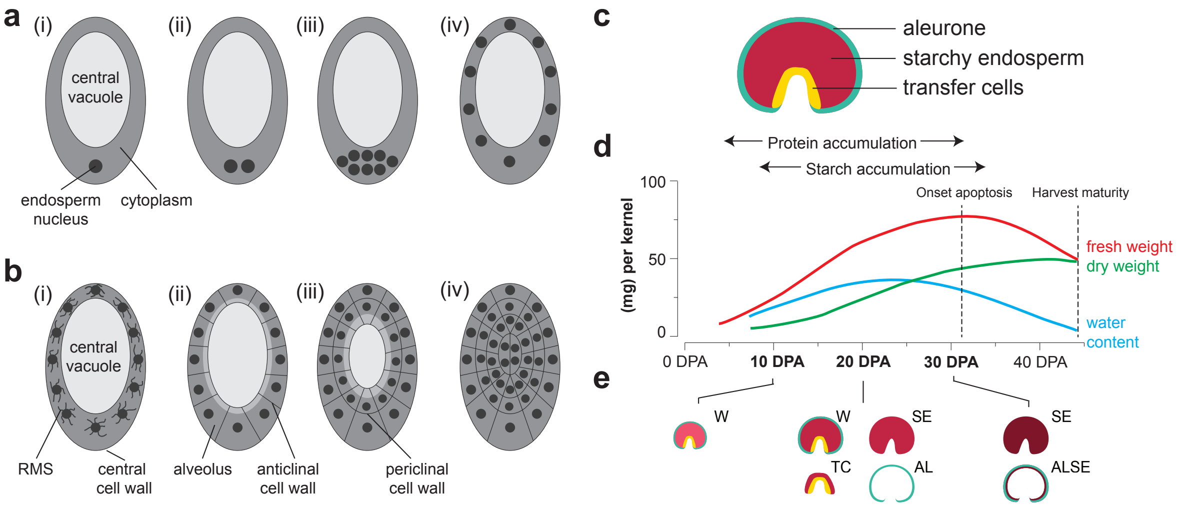


Fig. 5.1. Structure and developmental stages of the nuclear endosperm of cereals.
a, Early development of the endosperm coenocyte in cereals. (i) The triploid endosperm nucleus is located in the basal cytoplasm of the central cell, which encloses the central vacuole that constitute the largest portion of the central cell. (ii) Division of the nucleus generate a multinucleate cell, the endopserm conocyte, in absence of interzonal phragmoplast and cell wall formation. (iii) Eight nuclei are located in a single plane after the third round of cell divisions. (iv) Daughter nuclei migrate to the cytoplasm surrounding the central vacuole in uniform distances. **b,** Cellularization of the endosperm coenocyte in cereals starts with the (i) formation of radial microtubule systems on all nuclei, which initiate cellularization. (ii) Microtubules of neighboring nuclei form cell walls and generate alveoli (tube-like structures surrounding each nucleus), which are open towards the central vacuole. (iii) Alveolus nuclei divide and periclinal cells separate the peripheral cell and the new alveolus. (iv) Cell division continues until the central vacuole is completely filled with cells. **c,** The three major cell types of the mature endosperm analysed in this study. **d,** The temporal profile of grain development and transition points in the accumulation of starch, protein and water. **e,** The sampled cell types and developmental stages (W: whole endosperm; SE: starchy endosperm; TC: transfer cells; AL: aleurone cells). [*Manually adapted on basis from of a schematic illustration in (217) (**a** and **b**) and data from (349) (**d**).*]

grain and mediate transport of nutrients and photosynthate (mainly sucrose, monosaccarides and amino acids) from vascular tissue of the maternal plant into the endosperm *(217,351)*. They differentiate early in the cellularization phase and are characterized by an increased plasma membrane surface and extensive cell wall ingrowth that is important for efficient nutrient exchange. Aleurone cells form a single cell layer surrounding the starchy endosperm in wheat, however, thickness vary among grasses (e.g. three cell layers in barley or several cell layers in rice) *(216,217,348)*. AL cells produce hydrolases, glucanases and proteinases to mobilize starch and storage proteins in the starchy endosperm during seed germination *(217)*. Mature AL cells are cuboidal, rich in lytic and protein-storage vacuoles, which contain globoid bodies (a crystalline matrix of phytin, protein and lipid) and protein-carbohydrate bodies surrounded by lipid droplets *(216)*. Starchy endosperm cells compose the largest fraction of the endosperm and mainly synthesize starch in a series of enzymatic activities from Sucrose, which is transported from the leaf source tissue to the endosperm *(352)*. The second major compound of SE cells are storage proteins, including prolamins and globolins, which contribute approximately half of total protein in mature cereal grains *(353)*. These proteins are responsible for the viscoelastic and cohesive properties of wheat dough, which are important for food processing and bread baking.

Cell type differentiation is completed between 12 to 15 DPA and the endosperm enters the maturation phase, in which cells mainly increase dry weight by accumulation of starch and storage compounds *(349)* (Fig. 5.1d). In the later stage of seed maturation, cell expansion and water accumulation decline and the piled up solid compounds replace the fluid contents of the endosperm kernel, a process which implies dehydration of wheat grains *(354)*. Except AL cells, all other endosperm cell types undergo programmed cell death by approximately 30 to 35 DPA, the final stage of endosperm maturation. Membrane disassembly, DNA fragmentation and chromatin condensation are triggered *(354,355)* and facilitate mobilizing nutrients to the germinating embryo through hydrolysis of the exposed starch reserves by various enzymes that are produced and released from AL cells.

5.2 Dissecting the transcriptome of wheat endosperm

The individual endosperm cell types are morphological and functional highly different *(217)*. This requires a genome-wide understanding of spatial gene activity at different development (temporal) stages to identify key gene targets for an improved efficiency of breeding programs. Although endosperm gene expression profiling has been done in several other species including *A. thaliana* *(356)*, maize *(357)* or barley *(358)*, the characterization of bread wheat grain development was limited in comprehensiveness *(359)* or restricted to a global prospective without distinguishing homoeologous transcripts *(346)*. Therefore, deep RNA-seq profiling was applied in this study to monitor gene expression for different endosperm cell types at three developmental stages, which reflected the entire progression of starch and storage protein accumulation (10, 20 and 30 DPA) (Figs. 5.1d and e). Embryos were removed and grains cut in slices for the isolation of aleurone cells, transfer cells and starchy endosperm by manual dissection under the microscope. At 10 DPA, the whole endosperm (further used sample identifier ”10 DPA W”) was sampled, because individual cell types can not be isolated at this early stage. At 20 DPA a reference sample of the whole endosperm (”20 DPA W”) was produced as well as starchy endosperm (”20 DPA SE”), aleurone cells (”20 DPA AL”) and transfer cells (”20 DPA TC”) individually sampled. Due to tight adherence of starchy endosperm to the transfer cell layer, the TC sample included a small proportion of sourrounding SE cells. At 30 DPA grains were dissected into starchy endosperm (”30 DPA SE”) and aleurone cells (”30 DPA ALSE”). In the latter sample, AL cells tightly adhered to the outermost SE cells causing this sample to contain slight contamination of SE. For each of the seven tested conditions two biological replicates were sampled from grains of bread wheat plants grown in two greenhouses resulting in a total of 28 samples (Section 2.3).

5.2.1 RNA-seq read mapping and filtering

A reference-based strategy was applied for the analysis of gene expression on basis of the RNA-seq data set obtained in this study and the wheat CSS assembly and gene annotation generated by the IWGSC (Chapter 4). The separation of homoeologous chromosomes in the reference genome sequence allowed discriminating between homoeologous transcripts by using a ”first best match”-strategy, in which the genome-of-origin for an individual RNA-seq read was defined by the best reported alignment. Therefore, I mapped the obtained RNA-seq reads against the repeat-masked version of the wheat CSS assembly by using the well-established Bowtie/TopHat pipeline *(174,186)* (Bowtie version 2.1.0, TopHat version 2.0.8). Only the highest scoring TopHat alignment(s) with a maximum of two mismatches were considered for each read (parameters: --read-mismatches 2 --segment-mismatches 1 --max-multihits 20 -r 0). Subsequently, to avoid biased expression estimates caused by spurious assignment of RNA-seq reads to the incorrect wheat genome, I further filtered all obtained RNA-seq read alignment considering nine alignment scenarios and the following rules:

a Alignments of uniquely mapped singletons (only one read of a pair mapped) were accepted.

b Alignments of ambiguously mapped singletons (only one read of a pair mapped) were discarded.

c Alignments of reads were accepted if both reads of pair were mapped unambiguous to the same contig.

d Alignments of reads were accepted if reads of pair of were mapped to different contigs of the same chromosome arm (i.e. within the same genome).

e Read pair alignment were discarded if the individual reads were mapped to contigs of different chromosome arms and genomes.

f All alignments of a read pair were discarded if both reads were mapped ambiguously.

g If one end of a read pair was mapped uniquely to contig X and the other read end was uniquely mapped on contig X as well as to other contigs, both read alignments on contig X were accepted. All other alignment combinations were discarded.

h If one end of a read pair was mapped uniquely to contig Y and the other read ambiguously but only once to a contig Z, which originated from the same chromosome arm, the alignments to contig Y and Z were accepted. All other alignments were discarded.

i If one read was mapped unique and the other ambiguously, but never on a contig on the same chromosome arm, all alignments were discarded.

Overall, at least one read was aligned against the CSS assembly for 691 mio read pairs (Fig. 5.2a). Both reads were aligned for more than two third of the pairs (70%), whilst only one read was mapped for 16% (singletons). Thereby, the proportion of singleton read alignments reflected approximately the loss of sequence information in the CSS assembly, which has been previously observed by direct comparison of wheat fl-cDNAs to raw genomic shotgun sequences (Fig. 4.9). This was also largely consistent with studies using comparable NGS-based genome resources *(191)*. The majority of aligned reads [1,023 mio (63%)] were uniquely located in the CSS assembly (Fig. 5.2b). Ambiguous mappings (i.e. multiple alignments positions) with equal TopHat alignment score were observed for 234 mio aligned reads (14%). Due to the high sequence similarity between homoeologous gene copies in the bread wheat genome, the observed fraction of multiple read alignments was significantly increased compared to transcriptome analyses in diploid Triticeae genomes in which only around 1% of reads were ambiguously mapped *(191)*. With respect to the used alignment parameters (maximum two mismatches per read allowed) and

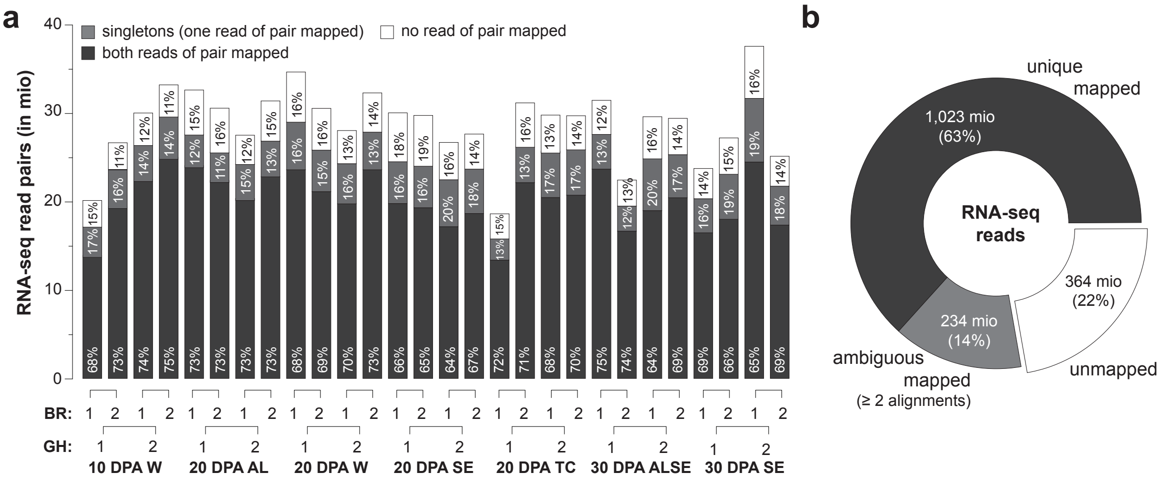
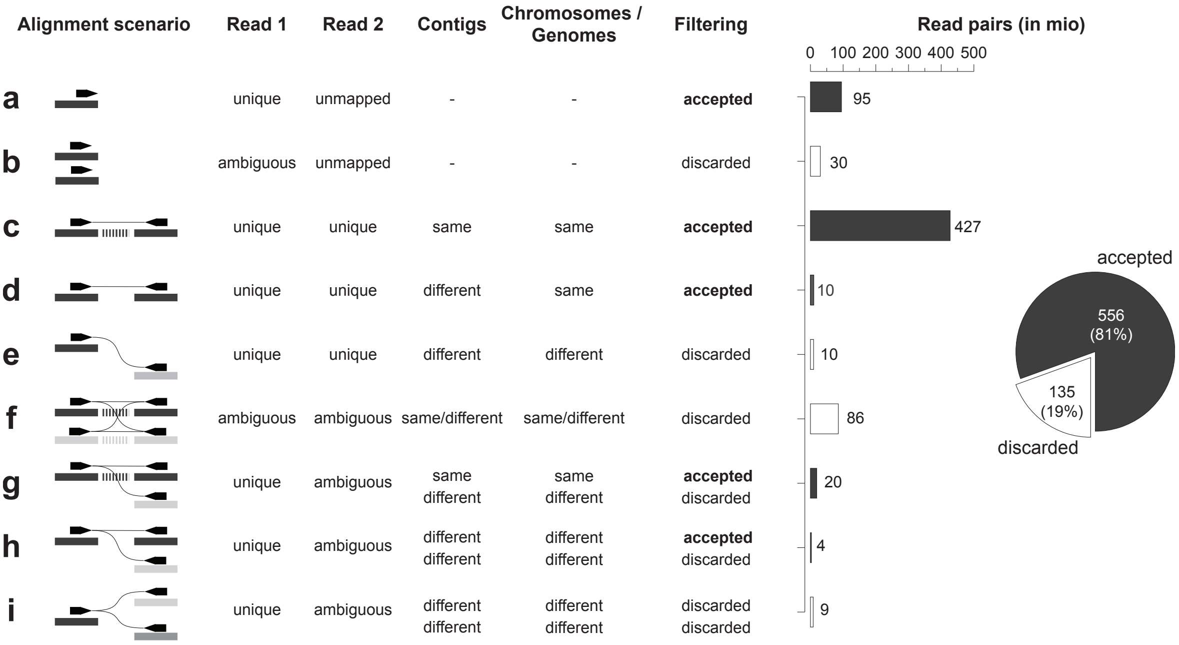


Fig. 5.2. RNA-seq mapping of individual endosperm samples to the wheat CSS assembly.
a, Number of RNA-seq read pairs of which both reads, one read or no read were aligned for each sampled RNA-seq library (GH: greenhouses; BR: biological replicates). **b,** Distribution of unique mapped reads (exact one mapping location), ambiguous mapped reads (multiple mapping locations with identical alignment score) and unmapped reads summarized for all samples.

a read length of 101 bp, the observed ratio between single and multiple mapped reads reflected between 98.4% to 99.2% sequence identity of homoeologs, in line with previous observations (Chapter 3) and other studies *(123,360)*.

The applied filtering rules resulted in a total of 556 mio accepted alignments (81%), while spurious alignments for 135 mio read pairs (19%) were discarded for the further analysis (Fig. 5.3). Ambiguously mapped singletons and read pairs contributed the majority of discarded alignments (86%). Contradictory read pair information, i.e. mapping of paired reads to different chromosome arms, was far less frequently observed.



Alignment scenario	Read 1	Read 2	Contigs	Chromosomes / Genomes	Filtering	Read pairs (in mio)
a	unique	unmapped	-	-	**accepted**	95
b	ambiguous	unmapped	-	-	discarded	30
c	unique	unique	same	same	**accepted**	427
d	unique	unique	different	same	**accepted**	10
e	unique	unique	different	different	discarded	10
f	ambiguous	ambiguous	same/different	same/different	discarded	86
g	unique	ambiguous	same different	same different	**accepted** discarded	20
h	unique	ambiguous	different different	different different	**accepted** discarded	4
i	unique	ambiguous	different different	different different	discarded discarded	9

Fig. 5.3. Classification of RNA-seq read pair mappings to nine alignment scenarios for stringent reads filtering.
Alignment of RNA-seq read pairs were categorized into nine groups (**a** to **i**) and filtered to reduce impact of spurious mapping of transcriptome sequences on the gene expression estimation. Contigs of the genome assembly are visualized by bold lines and the coloring depicts chromosome arm assignment. Reads are visualized as arrows and read-pairs connected by thin lines. The histogram shows the number of read pairs assigned to the corresponding alignment scenario. The pie chart shows the overall number of read pairs which were accepted and discarded for further analysis. See main text for further description of the individual alignment scenarios.

5.2.2 Refinement of the wheat gene annotation by incorporation of the endosperm transcriptome data

The IWGSC reference gene annotation (Chapter 4) provided the backbone for the transcriptome analysis conducted in this study. Although it has been previously suggested that the established gene annotation represented almost the entire gene catalogue of bread wheat (Section 4.2.2), the generated endosperm transcriptome data constituted an useful resource to refine the wheat gene annotation and to screen for additional genes as well as alternative splicing variants, which were completely absent or only lowly expressed in the IWGSC transcriptome resources.

For each of the seven endosperm samples the filtered RNA-seq alignments of the four corresponding samples (two BRs for two GHs) were merged and cufflinks *(306)* (version 2.0.2) applied to assemble these. The previously defined gene and transcript structures were supplied as reference annotation (parameter -g). A consensus gene set was generated with cuffcompare *(306)* (version 2.0.2), which clusters structures with identical intron boundaries into a non-redundant set of gene and transcripts. Then, the nucleotide sequences of novel assembled transcripts were extracted from the genome assembly and putative peptide sequences were predicted applying the OrfPredictor software *(222)* with sequence homology supported ORF selection against a combined set of proteins from *Brachypodium (42)*, rice *(45)*, sorghum *(40)*, maize *(41)* and *Arabidopsis (137)* [BLASTX *(227)* ($E \leq 10^{-5}$)]. All six reading frames were considered for transcripts located at previously unknown gene loci, whilst the strand direction was inferred from the IWGSC annotation for novel splicing variants located at already defined loci. Finally, the previously unknown gene loci were subjected to the same confidence class assignment that was applied for the IWGSC gene annotation on basis of protein-homology comparisons against high-quality gene sets of angiosperm genomes (Section 4.1.3).

This procedure identified a total of 401 novel high-confidence gene loci, of which five were classified as HC1 genes, i.e. were defined in full-length ($\geq$70% reference protein coverage) and most likely represented functional genes (Table 5.1). Seventy-seven novel genes were assigned to the medium gene confidence classes HC2 [16 genes (50%$\leq$ reference protein coverage $<$70%)] and HC3 [61 genes (30%$\leq$ reference protein coverage $<$50%)]. The majority fell into HC4, the lowest confidence class accumulating potential gene fragments and putative pseudogenes [319 genes ($<$30% reference protein coverage)]. On the contrary, a total of 15,625 additional splicing variants were detected including almost 12,000 transcripts for HC1 genes only. Thereby, no substantial increase was found in the total number of alternatively spliced genes (502 additionally detected AS genes). Functional enrichment analysis of the genes with novel

Table 5.1. Overview of the refined high-confidence gene set of bread wheat.

	IWGSC gene annotation					Refined gene annotation				
	HC1	HC2	HC3	HC4	Σ HC1-3	HC1	HC2	HC3	HC4	Σ HC1-3
Gene loci	55,249	14,367	15,475	39,110	85,091	55,254	14,383	15,536	39,429	85,173
Single exon	9,181 (17%)	3,230 (22%)	4,906 (32%)	20,375 (52%)	17,317 (20%)	9,160 (17%)	3,237 (23%)	4,937 (32%)	20,578 (52%)	17,334 (20%)
Multi exon	46,068 (83%)	11,137 (78%)	10,569 (68%)	18,735 (48%)	67,774 (80%)	46,094 (83%)	11,146 (77%)	10,599 (68%)	18,851 (48%)	67,839 (80%)
AS[a]	38,059 (69%)	7,916 (55%)	6,465 (42%)	8,728 (22%)	52,440 (62%)	38,413 (70%)	8,016 (56%)	6,513 (42%)	8,664 (22%)	52,942 (62%)
Transcripts	194,624	37,116	31,957	61,450	263,697	206,601	38,472	32,494	62,205	277,567
Exons[b]	538,250	94,864	74,630	117,530	707,744	550,031	96,383	75,273	118,376	721,687

[a] Alternatively spliced gene loci

[b] Exons of two or more transcripts were counted once if they have identical start and stop positions

splicing forms showed that these encoded for various functions including processes involved in endosperm development like glutamine biosynthesis *(361)* or sucrose metabolism *(362)*, but also for more basal cellular functions like chromosome organization, protein localization and protein folding (Fig. 5.4).

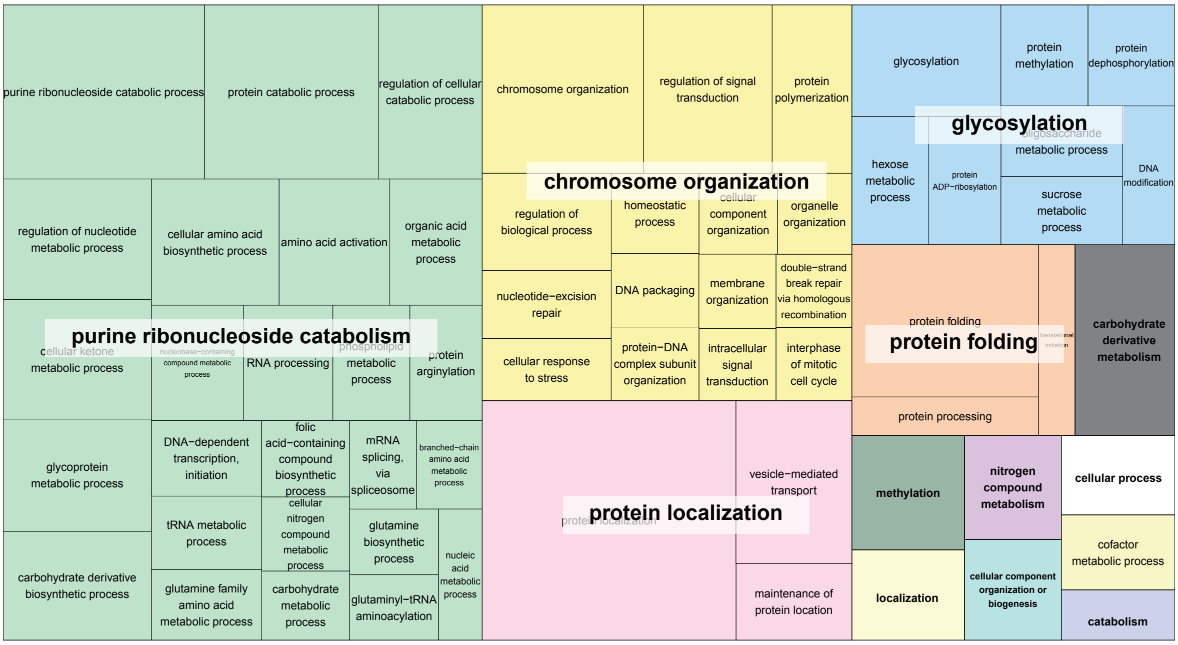


Fig. 5.4. Gene ontology categories analysis of genes with novel alternative splicing variants in the endosperm transcriptome.
Significant over-represented biological processes (*P* <0.01) were determined for high-confidence genes with novel predicted transcripts. Grouping of enriched gene ontology categories was generated by using the REViGO web server *(363)*. Box sizes correspond to significance of GO enrichment (*P* value).

5.2.3 Reproducibility of expression measures

To evaluate reproducibility of the expression measurements, the generated data set was tested for technical and biological variation. Therefore, the expression levels were determined for each individual replicate in "Fragments Per Kilobase of transcript per Million mapped reads" (FPKM) *(180)* by using cufflinks *(306)* (parameters: –G wheat-HC-gene-annotation.gtf -b wheat-reference.fa; version 2.0.2). Subsequently, the pairwise Pearson's correlation coefficients among replicates of a sample computed for the $\log_2$(FPKM+1)-transformed expression estimates.

The Pearson's correlation coefficients between biological replicates from plants grown in the same greenhouse ranged from 0.9078 for 30 DPA SE (greenhouse 2) to 0.9541 for 20 DPA AL (greenhouse 1) (Table 5.2). Values above 0.95 were observed between the two technical replicates generated for sample 20 DPA AL (biological replicate 1 for plants grown in greenhouse 1). With the exception of sample 30 DPA ALSE (R^2 of 0.80), slightly decreased but still good agreement of gene expression was found among all samples (R^2 >0.89) in pairwise comparisons

Table 5.2. Pearson's correlation coefficient (R^2) of gene expression levels estimated for biological replicates grown in the same and in different greenhouses.

Sample	R^2 within greenhouse 1	R^2 within greenhouse 2	Mean R^2 of pairwise comparisons of replicates between greenhouses
10 DPA W	0.9249	0.9285	0.9110 ± 0.0178
20 DPA AL	0.9541	0.9263	0.9053 ± 0.1110
20 DPA W	0.9399	0.9242	0.8717 ± 0.0026
20 DPA SE	0.9252	0.9125	0.8926 ± 0.0313
20 DPA TC	0.9367	0.9340	0.9018 ± 0.0021
30 DPA ALSE	0.9182	0.9229	0.8033 ± 0.0265
30 DPA SE	0.9163	0.9078	0.8991 ± 0.0043

between greenhouses.

The generally good agreement between biological and technical replicates were also reflected in a principle component analysis of gene expression across samples (Fig. 5.5). However, an unexpected sample clustering was evident for samples 20 DPA W (greenhouse 2) and 30 DPA ALSE (greenhouse 2), which corresponded to the low correlation coefficients for 30 DPA ALSE between greenhouses and indicated a potential swap of labels during the experimental sample preparation before sequencing. Therefore, these two replicates were excluded from the subsequent analysis. Although smaller-scale variation were found in the gene expression measurements for samples of plants grown in different greenhouses, overall high agreement of gene expression for technical replicates and biological replicates confirmed accuracy of RNA-seq expression quantification and demonstrated the high reproducibility of the conducted experiments *(364)*.

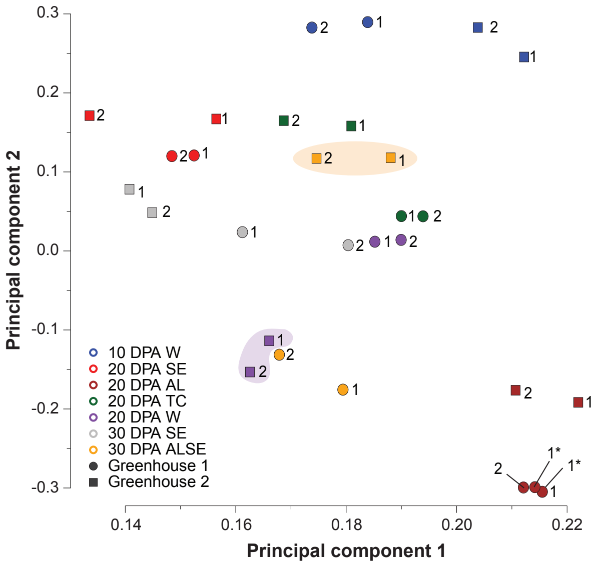


Fig. 5.5. First and second principal component of gene expression among replicates.
Numbering right to each data point represents replicate number and stars mark technical replicates. Highlighted samples [20 DPA W (greenhouse 2) and 30 DPA ALSE (greenhouse 2)] were excluded for further analysis due to potential swapped labels during sample preparation.

5.2.4 *In silico* validation of gene expression measurements

The high similarity of coding-sequences among homoeologous genes *(123)* might be problematic for determining the genome-of-origin of RNA-seq short reads obtained for the hexaploid wheat transcriptome and, thus, may impact on accurate measurement of gene expression levels. Therefore, I performed an *in silico* evaluation experiment to validate the accuracy of the computed expression levels and to confirm the applicability and reliability of the implemented methods (Fig. 5.6a). Illumina-like artificial short read pairs, which represented an experimental setup that is comparable to the real data set, were generated on basis of the annotated transcript structures from the CSS assembly by using FluxSimulator *(365)* (parameters: 101 bp read length, 200 bp paired-end insertion size, Illumina read error model and random gene expression levels; version 1.2). The simulated sequencing reads were aligned against the reference genome sequence and filtered applying the same protocol as described in Section 5.2.1. Gene expression level were estimated in FPKM with cufflinks *(182)* (version 2.0.2). Polynomial regression fits of the simulated and estimated $\log_2$(FPKM+1)-transformed expression values were computed for the entire gene set as well as for a total of 19,728 genes, which formed single-copy homoeologous gene triplets (Sections 4.3.3 and 5.4.3) using the `loess.smooth`-function implemented in R (parameter: span=0.2).

Overall, 16.5 mio out of 17.5 mio simulated RNA-seq read pairs were successfully aligned against the bread wheat genome assembly (94%). The read filtering step removed a comparable

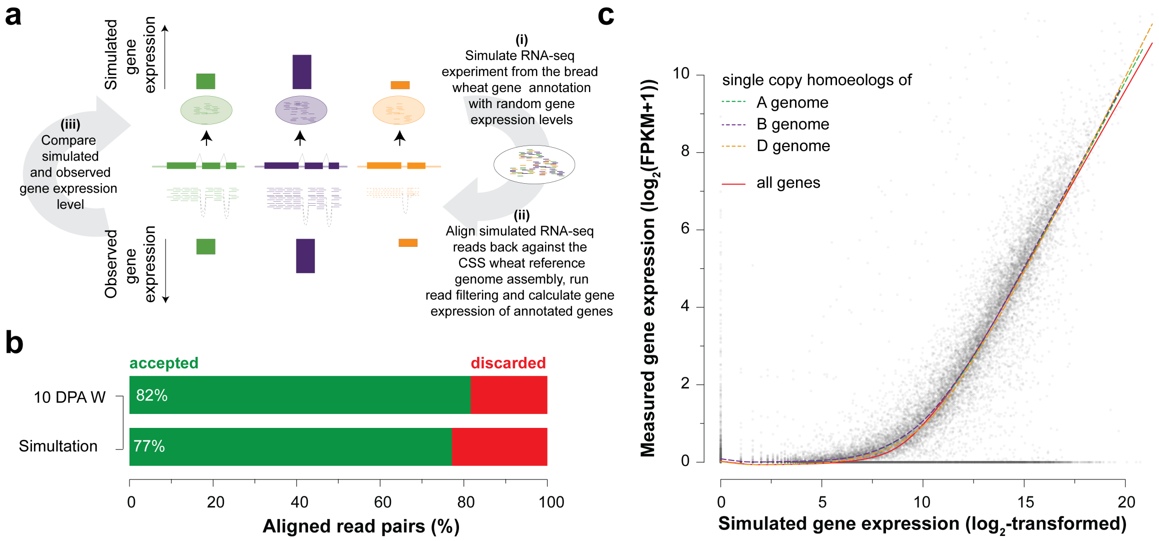


Fig. 5.6. Validation of homoeologous gene expression measurements.
a, Workflow for the validation of gene expression measurements in a polyploid context with a RNA-seq simulation experiment. **b,** Fraction of aligned read pairs which are accepted and discarded in the filtering step. **c,** Comparison of simulated and measured gene expression levels. Dots show single measurements and lines represent a polynomial fit of the expression measurements for all genes (red solid line) for homoeologous genes forming single-copy gene triplets (dashed lines).

number of simulated reads as observed for real expression data (Fig. 5.6b). Less than 1% of RNA-seq reads were aligned to a wrong contig in the CSS assembly after the filtering step. Although expression was underestimated for low abundant genes, a generally good correlation was observed between the simulated and measured expression levels (Fig. 5.6c). Importantly, the good agreement held also true for single-copy homoeologous genes, which confirmed the correctness in calculating genome-specific expression levels. In summary, these observations suggested a high reliability of the computed expression values, which was essential for excluding any technical bias in the subsequent expression analysis.

5.2.5 Computation of gene expression and differential expression tests

Excluding the likely swapped samples 30 DPA W (greenhouse 2) and 30 DPA ALSE (greenhouse 2) (Fig. 5.5), the expression levels of wheat high-confidence genes were calculated in FPKM *(180)* and tested for significant differences in pairwise comparisons between samples. Therefore, cuffdiff *(182,306)* was used (parameters: –N –b wheat-reference.fa, version 2.0.2), which converts the alignments of RNA-seq reads into models of fragment counts combined with an estimate of uncertainty in biological variation. All subsequent analysis were restricted on 85,173 high-confidence genes, which have been classified into levels HC1 to HC3 (Table 5.1). Wheat genes of the HC4 gene set were not considered in this study as this class most likely included many (deteriorated) gene fragments and pseudogenes (Section 4.3).

Considering all genes with FPKM greater than zero would overestimate presence and absence of gene expression in the qualitative analysis of gene expression. Therefore, a lower expression limit of 0.02 FPKM was defined based on the mean 10th percentile of the calculated gene expression levels across all endosperm samples (Table 5.3). Similar to expression estimation using microarry technology, all further statistical testing expression values were $\log_2$(FPKM+1)-transformed to decouple the signal intensity from random error *(364,366)*, whereupon the addition of 1 to all estimated FPKM values avoids negative values after the $\log_2$ transformation.

Table 5.3. Gene expression level statistics for high-confidence wheat genes (HC1-3) for individual endosperm samples.

Sample	Gene expression level (FPKM)						Expressed	Expressed
	Mean	Median	5th	10th	90th	95th	genes[(a)]	transcripts[(a)]
10 DPA W	2.55 ± 44.39	0.17	0.01	0.02	1.92	4.58	37,046 (44%)	57,486 (49%)
20 DPA AL	1.91 ± 29.87	0.17	0.01	0.02	1.88	4.19	37,381 (44%)	59,880 (51%)
20 DPA W	2.29 ± 50.20	0.16	0.02	0.02	1.61	3.62	35,153 (41%)	51,786 (44%)
20 DPA SE	2.92 ± 72.07	0.17	0.02	0.03	1.66	3.93	35,097 (41%)	51,721 (44%)
20 DPA TC	2.18 ± 56.39	0.12	0.01	0.01	1.30	2.92	37,384 (44%)	56,017 (48%)
30 DPA ALSE	1.77 ± 23.84	0.19	0.02	0.03	1.74	3.85	34,588 (41%)	52,487 (45%)
30 DPA SE	2.38 ± 55.58	0.15	0.01	0.02	1.52	3.51	35,736 (42%)	53,741 (46%)
Overall	2.28 ± 47.48	0.16	0.01	0.02	1.66	3.80	46,487 (55%)	117,620 (43%)

[(a)] Only genes and transcripts with minimum expression level FPKM≥0.02 (10th percentile of overall gene expression) were considered to be expressed.

5.3 The global transcriptional landscape of bread wheat endosperm

5.3.1 Quantitative analysis of gene and transcript expression

Overall, 46,487 out of 85,173 high-confidence genes (55%) and 117,620 out of 277,567 transcripts (43%) were detected in the RNA-seq data set (Fig. 5.7a and Table 5.3). Thereby, the three wheat genomes contributed about equally to the number of expressed genes and transcripts in the endosperm as a whole (18% to 19% of genes and 14% to 15% of transcripts) as well as in individual cell types and developmental stages (Fig. 5.7b). Strikingly, significant differences were present in the spatiotemporal expression distribution of genes (i.e. sum of all transcript variants at a certain locus) compared to that of individual alternative splicing variants (Fig. 5.7c and d). Whilst more than half of the expressed genes were detected in all sampled endosperm cell types and time points, only 14% of transcripts were so. On the contrary, 10% of genes were found to be specifically transcribed in a single condition, whereas one quarter of transcripts were detected in a single sample. Again, no significant differences among the three genomes were evident.

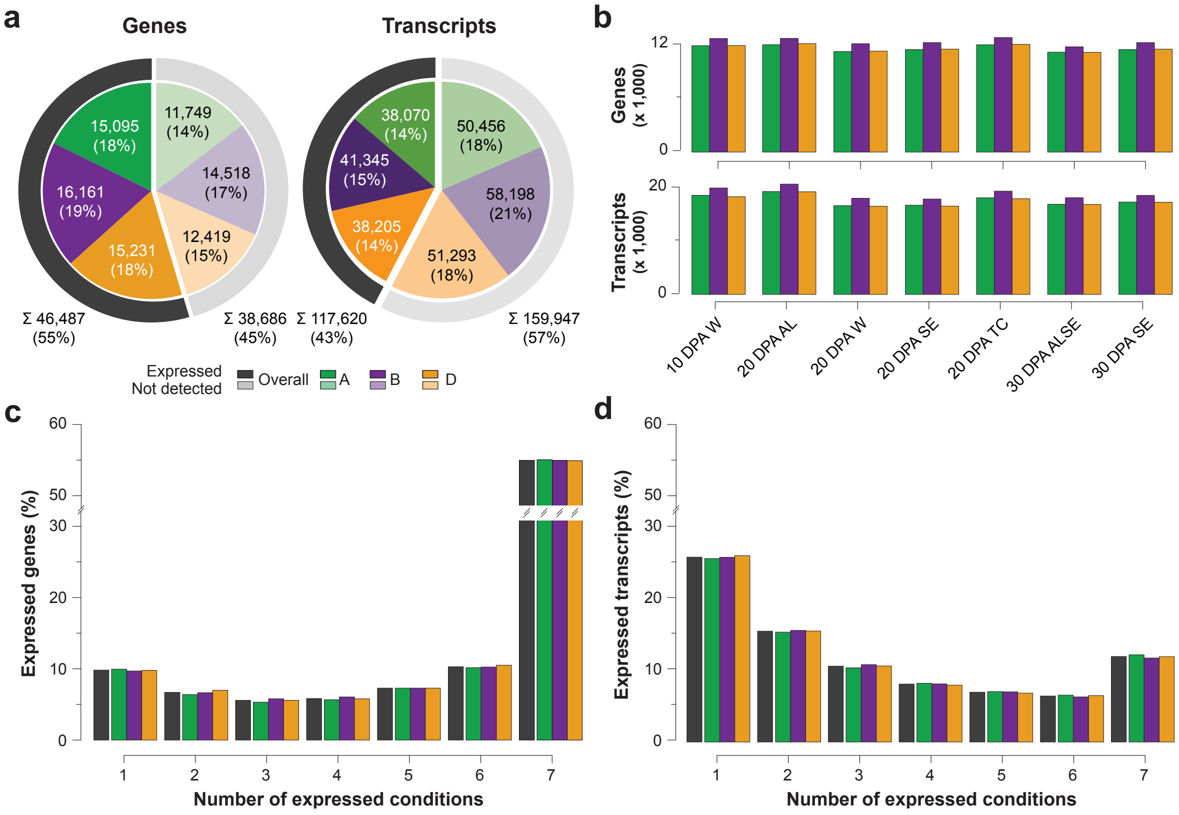


Fig. 5.7. Distribution of endosperm gene and transcript expression across the A, B and D genomes. Number of expressed high-confidence wheat genes and transcripts **a,** across all and **b,** in individual endosperm cell types and developmental stages. Number of samples in which **c,** genes and **d,** transcripts were observed to be expressed.

5.3.2 Identification of preferentially expressed genes

Genes that are expressed at a higher level under a certain spatiotemporal condition might constitute key regulators and marker genes *(367)* and, therefore, are interesting targets for the improvement of grain quality attributes. Here, such "preferentially expressed genes" (PEGs) were identified for whole grain (W) at 10 DPA and individual cell types at 20 DPA (AL, SE and TC) as well as 30 DPA (AL and SE) by using two complementary approaches[(1)]. First, candidate PEGs were defined on basis of non-overlapping 95% confidence intervals of gene expression between each tested condition and a corresponding reference group formed by the remaining samples (Table 5.4). Thereby, the lowest CI of gene expression in the tested condition had to be larger than the highest CI in the reference group. Secondly, differential expression analysis was performed between conditions and reference groups with cuffdiff *(182)*. Genes with significant higher gene expression in the tested condition were selected as candidate PEGs [false discovery rate (FDR) <0.05]. Finally, the two sets were merged and candidate genes identified by either of these two approaches defined as the final set of PEGs.

Across genomes a comparable low total number of PEGs was observed (Table 5.4). Between individual cell types and developmental stages the number of identified PEGs varied considerably ranging from 136 PEGs in 20 DPA TC to 644 PEGs in 20 DPA AL. As revealed by a functional enrichment analysis, the determined PEGs encoded for proteins with annotated gene ontology categories that well agreed with the observed transcriptional activity and the known functional characteristics of the sampled cell types (Table A.1). For example, proteins function in carbohydrate metabolic processes and glycolysis were enriched in the set of PEGs in 10 DPA W *(346,358)*, whilst lipid metabolism, structural development, carbohydrate metabolic processes and amino acid biosynthesis were predominantly detected for AL-specific PEGs *(346)*, carbohydrate and saccharide metabolism for SE-specific *(368)* or proteolysis and defense response genes for TC-specific PEGs *(369)*, respectively.

Table 5.4. Identification of preferentially expressed genes for individual endosperm cell types.

Condition[(a)]	Reference group	Number of PEGs			
		A	B	D	Σ
10 DPA W	all other samples	108	106	100	314
20 DPA AL	20 DPA SE, 20 DPA TC	197	232	215	644
20 DPA TC	20 DPA AL, 20 DPA SE	52	49	35	136
20 DPA SE	20 DPA AL, 20 DPA TC	30	24	29	83
30 DPA AL	30 DPA SE	136	153	141	430
30 DPA SE	30 DPA ALSE	83	84	76	243

[(a)] 20 DPA W was not tested for PEGs, because all cell types were present as individual samples.

[(1)] I gratefully acknowledge Karl Kugler for the definition of preferentially expressed genes and further gene ontology enrichment analysis.

5.3.3 Spatiotemporal differences in gene expression

The morphological and functional differences of the analysed endosperm cell types (Section 5.1) were clearly apparent from hierarchical cluster analysis of the whole endosperm transcriptome (R package pvclust *(318)* with Pearson's correlation distance; average linkage clustering and 1,000 times bootstrap re-sampling) (Fig. 5.8). Aleurone cells (samples 20 DPA AL and 30 DPA ALSE) formed an expression cluster, which was separated from whole grain (samples 10 DPA W and 20 DPA W), clean starchy endosperm cells (samples 20 DPA SE and 30 DPA SE) and transfer cells (20 DPA TC). In the later group gene expression similarities related to the developmental stages seemed to triumph over cell type similarities. Samples from the 20 DPA stage clustered together, while clean starchy endosperm and samples containing transfer cells did not. Although transfer cells and starchy endosperm cells are functionally highly different *(216,217)*, the transfer cell sample clustered with samples incorporating mainly starchy endosperm cells. However, low bootstrapping probabilities indicated uncertain placing of 20 DPA TC in the clustering dendrogram, which was most likely due to tight adherence of starchy endosperm cells that hampered the pure dissection of transfer cells (Section 5.2).

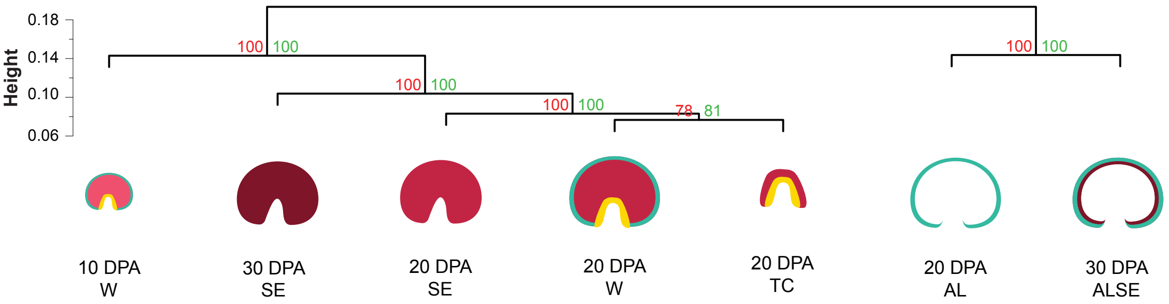
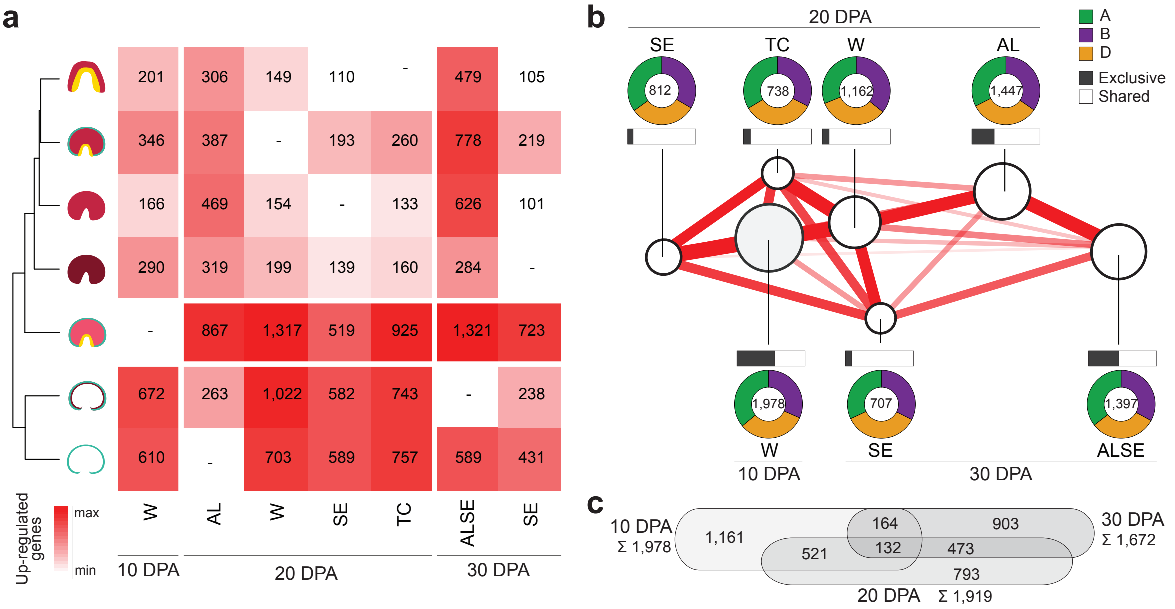


Fig. 5.8. Spatiotemporal hierarchical cluster analysis of endosperm gene expression.
Similarity in gene expression among different cell types and developmental stages was investigated by hierarchical cluster analysis using Pearson's correlation distance average linkage clustering. Significance estimates were determined and are shown for each branch. Approximated unbiased *P* values were calculated by multiscale bootstrapping (green numbers) and bootstrap probabilities calculated by normal bootstrapping (red numbers).

5.3.4 Qualitative analysis of differential gene expression

The known functional differences between cell types were also apparent from analysis of differential gene expression regulation. A total of 4,384 differentially expressed (DE) genes (9%) were identified in pairwise comparisons between the seven samples. Consistent with strong separation of aleurone cells and starchy endosperm in the hierarchical cluster analysis discussed above, a high number of DE genes were found between these two cell types (e.g. 1,058 DE genes between 20 DPA AL and 20 DPA SE) (Fig. 5.9a). Furthermore, a high number genes were differentially regulated between time points with 1,993 genes between the most distant time points early differentiation (10 DPA W) and maturation (30 DPA ALSE), 1,800 genes between the neighbouring phases late differentiation (20 DPA W) and maturation (30 DPA ALSE) and 1,663 between the two differentiation stages (10 DPA W and 20 DPA W), respectively. Strong initial gene expression was found in whole endosperm at 10 DPA, which had with 1,978 the highest number of up-regulated[2] genes compared to any other sample. Only 707 up-regulated genes were found in mature starchy endosperm cells (30 DPA SE), which was consistent with decrease in gene expression and initiated cell death in these cells *(370)*. Contrary, transcription was continuing in aleurone cells (1,397 up-regulated genes in 30 DPA ALSE), which actively participate in nutrition of the growing embryo in later time points *(354,355)*.

To further investigate the similarities in modulated gene expression across cell types and developmental phases, a spatiotemporal analysis was performed for the identified significant up-regulated genes. Therefore, I constructed a network, in which a node was created for each sample (Fig. 5.9b). To represent the transcriptional relationships between the seven tested conditions, I connected the nodes by undirected edges, which were weighted by the number of commonly



	W	AL	W	SE	TC	ALSE	SE
	201	306	149	110	-	479	105
	346	387	-	193	260	778	219
	166	469	154	-	133	626	101
	290	319	199	139	160	284	-
	-	867	1,317	519	925	1,321	723
	672	263	1,022	582	743	-	238
	610	-	703	589	757	589	431

Fig. 5.9. Spatiotemporal analysis of differentially gene expression.
a, Number of significant up-regulated HC wheat genes in pairwise comparison of sampled endosperm cell types and time points (up-regulated in sample row compared to sample column). **b,** Network representation of interrelation of endosperm samples (nodes) in commonly up-regulated genes. Node sizes correlate with the total number of significant up-regulated genes and edge width with the number of shared up-regulated genes. Circle diagram visualize distribution among wheat genomes. The fraction of sample-specific up-regulated genes is depicted by the filled parts of the bars. **c,** Number of up-regulated genes shared between the three sampled developmental stages. Respective samples of individual cell types were merged for 20 DPA and 30 DPA.

[2] In the following, the term ”up-regulated” was used to denote the sample with higher gene expression as representative for the direction of the DE test. In some cases this might not be correct as it could not be decidable if the higher expressed genes has been enhanced or the lower expressed genes repressed.

up-regulated genes. Then, the network was topologically arranged on basis of the edge weights by using the ”edge-weighted force directed” layout algorithm implemented in Cytoscape *(371)* (version 3.0.2). This strategy placed nodes with a higher number of shared genes closer to each other.

Consistent with the previous observations, aleurone cells were separated from whole grain and samples containing mostly starchy endosperm. The majority of up-regulated genes were either shared with the mixed samples or between clean samples of the same cell type. Thereby, the three homoeologous wheat genomes contributed about equally to the number of differentially expressed genes. When comparing gene expression across the different developmental stages, approximately half of the up-regulated genes were found to be exclusively up-regulated for each phase (Fig. 5.9c). The largest overlaps were observed between adjacent time points (521 commonly up-regulated genes in 10 and 20 DPA and 474 genes in 20 and 30 DPA, respectively), while only 161 genes were commonly up-regulated for 10 and 30 DPA. This underpinned the on-going, partially overlapping functional specification of endosperm cells on the developmental course to maturation *(216)*. Although a large number of genes were shared for aleurone cells sampled at 20 DPA and at 30 DPA, noteworthy, for both time points a considerable amount of genes were found to be exclusively up-regulated genes. This observation suggested substantial transcriptional differences between aleurone cells in the differentiation phase and maturation phase. With respect to the molecular functions determined for corresponding PEGs, the observed patterns indicated a functional change of aleurone cells to produce other proteins and enzymes in mature seeds that are more involved in transmembrane transport and embryo nutrition *(216)*.

5.4 Exploring co-regulation of gene expression in wheat endosperm

Identification of genes that exhibit common transcriptional activity under certain functional constellations is essential for the further analysis of the regulatory mechanisms within biological systems *(372)*. Grouping of those ”co-expressed” genes into clusters allows characterizing the complex interactions that concert gene expression and facilitates unravelling functional relationships between genes. This study aimed at contributing to an understanding of bread wheat endosperm development by the identification of well-defined co-expression clusters. Such groups of genes constitute the starting point to further screen in detail for gene-to-gene associations and for the regulatory mechanisms underlying common transcriptional activity of genes.

5.4.1 Identification of endosperm co-expression clusters

One method to group co-expressed genes is k-means clustering *(373)*, which separates the dimension space into a pre-defined number of clusters (k). This approach aims at computing the most compact clustering with closest distances between commonly grouped genes by minimizing the within-cluster sum of squares. In this study, I utilized the k-means algorithm implemented in the R package amap *(374,375)* with Pearson's correlation distance. As the resulting k-means clustering is largely dependent on the chosen initial number of clusters, the most appropriate value was selected by testing different parametrizations in seven independent clustering rounds ($k \in [6, 13]$). The quality of each resulting clustering was evaluated on basis of the silhouette coefficient *(376)*, a numerical value measuring the discriminative power in the gene-to-cluster assignments, by using the `silhouette`-function implemented in the R package cluster *(377)*. Large silhouette values (almost 1) suggest strong clustering, while small values (around 0) indicate that data points fall between two clusters and negative silhouette values are associated with uncertain clustering of observations. For each iteration I determined the mean silhouette coefficient over all clusters, thereby discarding poorly-defined clusters with negative silhouette coefficient (Fig. 5.10a).

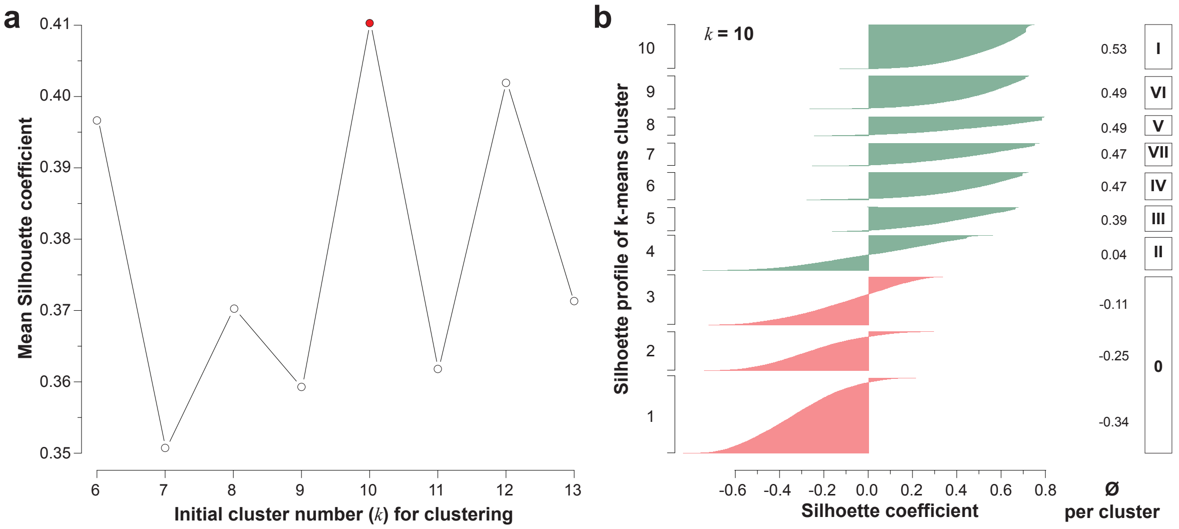


Fig. 5.10. Selection of the cluster sizes and silhouette plot for co-expression clustering.
a, The k-means clustering was repeated for different initial cluster numbers and the mean silhouette coefficient of all well-defined cluster calculated. Maximum silhouette value was achieved for k-means clustering with a k of 10 (red). **b,** Distributions of the silhouette coefficients obtained for each of the identified co-expression clusters. Clusters with positive mean (⌀) coefficients are colored green (Clusters I to VII), whereas clusters with negative mean coefficients are colored red (Cluster 0).

In this study an initial k of 10 was chosen as most appropriate clustering of the transcriptome data resulting in seven well-defined co-expression clusters with positive silhouette coefficients, i.e. good separation in the dimension space and stable cluster assignments (Clusters I to VII) (Fig. 5.10b). The remaining three clusters with negative silhouette coefficients, which indicated uncertain cluster assignments, were combined into a ”zero” cluster (Cluster 0) for the further analysis.

The seven clusters with positive silhouette coefficients (Clusters I to VII) contained between 2,257 to 5,369 genes (24,826 genes in total) and showed clear and distinct gene expression profiles with preferential transcription in a subset of the tested spatiotemporal conditions (Fig. 5.11a). On the contrary, balanced gene expression was observed across all endosperm cell types and developmental stages for Cluster 0. This was consistent with negative silhouette coefficients, which already indicated poor between-cluster separation for the included genes. Interestingly, all clusters contained a similar number of genes from the A, B and D genome as well as similar overall expression profiles.

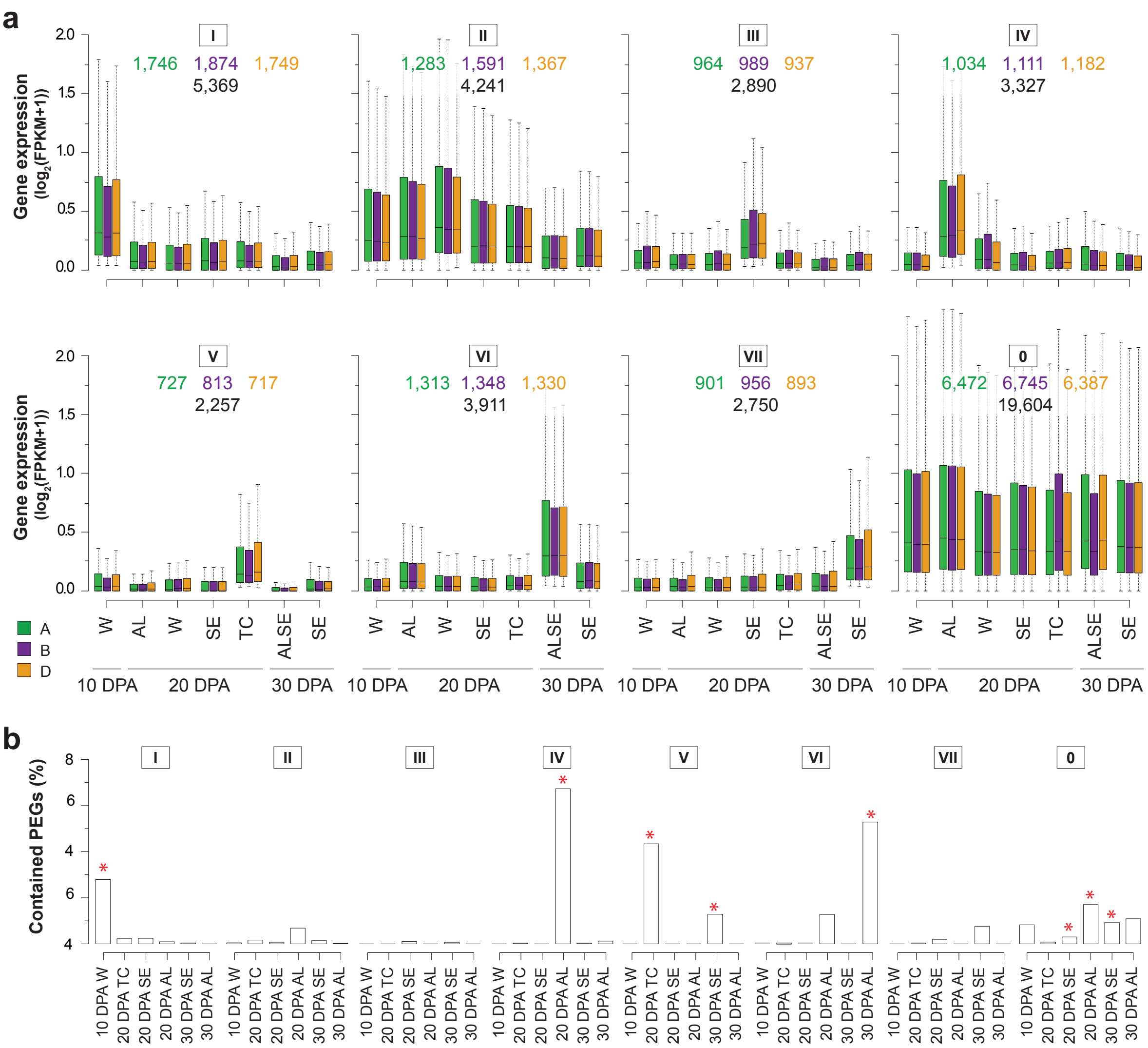


Fig. 5.11. Gene expression profiles of co-expression clusters.
a, For each co-expression total the boxplots visualize the gene expression levels across the seven spatiotemporal endosperm samples. Numbering counts clustered genes in total in for individual genomes. **b,** Number of preferentially expressed genes that were assigned to the individual co-expression clusters. Red stars indicate a significant enrichment of a cluster for PEGs (Pearson's chi-squared test with Bonferroni adjusted *P* value <0.05).

5.4.2 Functional characterization of the identified co-expression clusters

To investigate the biological meaning of commonly grouped genes, each co-expression cluster was tested for significant enrichment of PEGs (Fig. 5.11b) and over-represented gene ontology categories (Table A.2). As further discussed below, this analysis revealed distinct functionally characteristics for each cluster, which accompanied the observed spatiotemporal gene expression profiles found for commonly grouped genes.

Early endosperm differentiation (10 DPA)
Cluster I represented the early developmental phase when cell divisions are still occurring in the periphery of the endosperm and the transcription of storage proteins and accumulation of starch have been initiated. Genes were found predominantly expressed in 10 DPA W and a significant proportion of PEGs for early development were included in this cluster. Consistent with proteomic studies of wheat endosperm development wheat *(368)*, this cluster was enriched for genes encoding various catabolic and metabolic processes like sucrose metabolism, glucose metabolism, carbohydrate metabolism and nitrogen metabolism as well as proteolysis, signaling and cellular component organization.

Endosperm differentiation (10 DPA to 20 DPA)
Clusters II to V grouped genes expressed predominantly during the endosperm differentiation phase, where the accumulation of storage protein and starch accumulation reaches maximum (Fig. 5.1d). Cluster II showed cell type-unspecific expression profiles, but connected the early and intermediate differentiation phases, whereas gene expression profiles Clusters III to V characterized particular endosperm cell types at 20 DPA. Genes in Cluster III were mainly expressed in starchy endosperm cells (20 DPA SE) and encoded for proteins involved in accumulation of carbohydrate and storage compounds including cellular macromolecule metabolic process, monosaccharide metabolic process or glutamine family amino acid metabolic process. Cluster IV exhibited increased expression in 20 DPA AL, included a significant proportion of genes preferentially expressed in 20 DPA AL and was enriched for processes related to catalytic activity, lipid metabolic processes and carbohydrate metabolism. Cluster V grouped genes expressed primarily in transfer cells, which are involved in transport (e.g. anion transport and drug transmembrane transport), are responsive to stimuli (e.g. response to light or water stimulus) and are related to defense-like proteins *(369)*.

Endosperm maturation (30 DPA)
The remaining clusters showed increased expression levels in mature wheat endosperm cells. For Cluster VI primarily gene expression was observed for aleurone cells at 30 DPA and significant enrichment for the corresponding PEGs found. Lipid, vitamin and amino acid metabolism as well as cellular response and transmembrane transport activity were significant over-represented gene ontology categories. Cluster VII grouped genes, which were found to be expressed in mature

starchy endosperm cells (30 DPA SE) and which encode mainly for proteins that negatively regulate various cellular processes including translation, protein metabolism, macromolecule biosynthesis. Moreover, proteins encoding for signalling and stress response as well as acting in autolysis and programmed cell death like chitin catabolism *(378)* were significantly over-represented in Cluster VII.

5.4.3 Gene expression regulation of homoeologous genes

So far, no indications for genome asymmetry and transcriptional differences among the wheat genomes were evident, neither in terms of number of genes nor in gene expression level. Therefore, the congruences in the gene expression profiles and the co-regulation of single-copy homoeologous genes were directly analysed. Following the same protocol as described in Section 4.3.3 of this thesis, a set of 6,576 homoeologous gene triplets [6,576 x 3 = 19,728 genes (HC1 to HC3)] was defined on basis of pairwise bi-directional protein BLAST *(227)* searches between the updated genes sets of the A, B and D genomes[3]. Although these triplets only constituted a snapshot of the entire wheat genome and discarded any genome dynamics (e.g. copy number variations) that might constitute an additional layer of complexity, focussing on single-copy homoeologs allowed measuring regulatory influences acting on genes, which have been retained during common polyploid evolution. However, the previous analysis showed that these homoeologous triplets were a good representation of the entire wheat genome (Section 4.3.3), which enabled making conclusions of the structural and functional impacts of polyploidization on a genome-wide level.

Global patterns of homoeologous gene expression divergence

Overall, at least one homoeologous gene was found to be expressed for 5,939 triplets (Fig. 5.12a). Among these, all three homoeologs were transcribed for 4,912 triplets (83%), while two homoeologs were expressed for 589 triplets (10%) and exclusively one homoeolog for 438 triplets (7%), respectively. The observed distribution clearly deviated to an assumption of complete independence of homoeologs gene expression regulation, which was tested with random assignment of wheat genes into triplets. Significantly more triplets then expected by chance were entirely silenced as well as more triplets completely retained expression for all homoeologs [1,000 permutations (P <0.05)]. Since the analysed genes were derived from a common ancestor, these findings reflected the relatedness of homoeologs and indicated maintained expression for genes forming triplets and contribution of each homoeologous gene copy to the entire wheat grain transcriptome.

Interestingly, while globally expression was maintained, significant differences in cluster assignments of homoeologous genes were evident (Fig. 5.12b). All members were placed into the same co-expression cluster for 28% of the triplets (1,663), while only two out of three homoe-

[3] I gratefully acknowledge Sapna Sharma for the computation of the homoeologous gene triplets on basis of the updated wheat gene annotation.

ologs were assigned to the same co-expression cluster for 41% of the triplets (2,416). For almost one third of the triplets all three homoeologous genes fell into separate clusters (1,860 triplets). Thereby, a uniform distribution across genome pairs was found for homoeologs placed into the same co-expression clusters (Tables A.3 to A.5). A total of 818 A- and B-genome encoded homoeologs (14%), 794 A- and D-genome encoded homoeologs (13%) and 804 B- and D-genome encoded homoeologs (14%) were placed in the same co-expression clusters. A balanced distribution was also observed for completely silenced homoeologs (1,150 genes of the A genome, 1,103 of the B genome and 1,123 of the D genome).

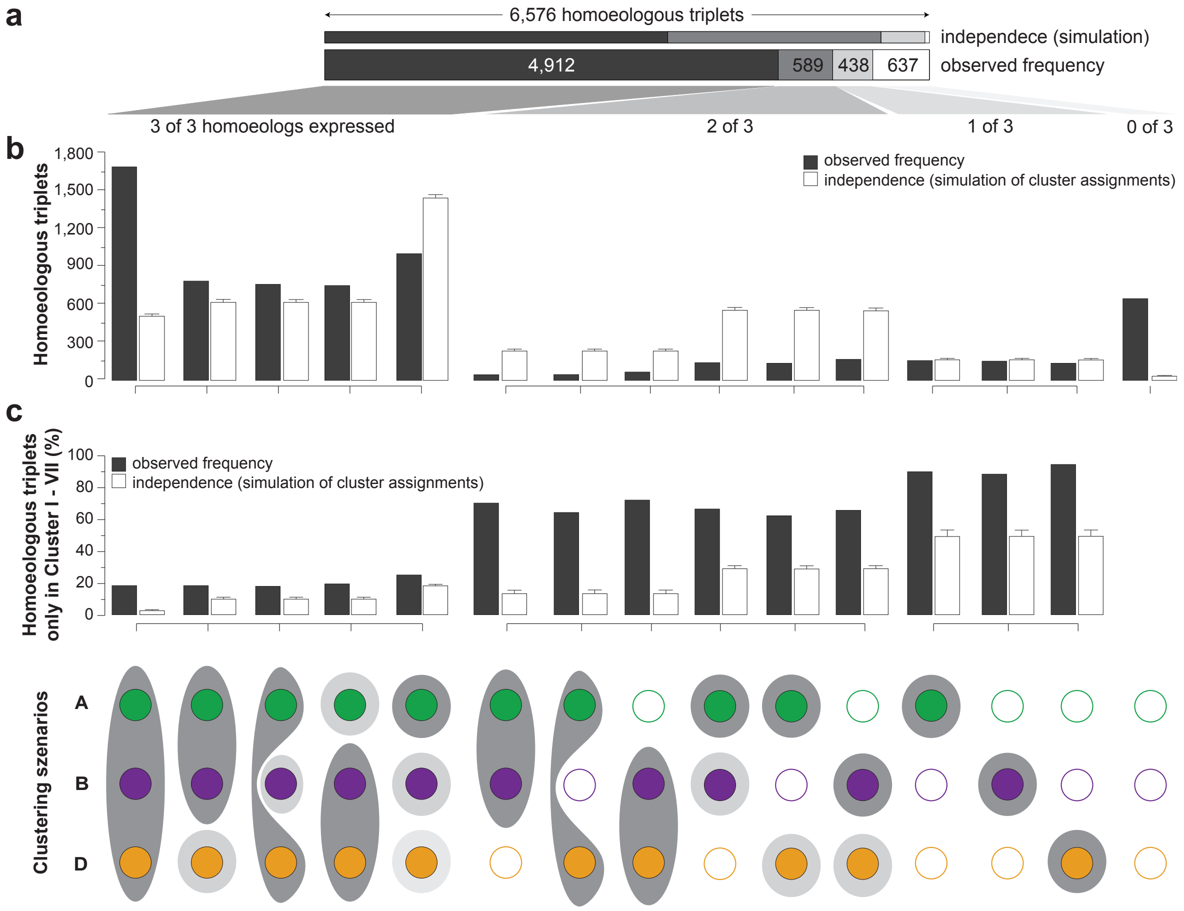


Fig. 5.12. Diverged co-expression cluster assignments for homoeologous gene triplets. All possible clustering scenarios for homoeologous gene triplets are visualized by the illustrations along the x-axis. Colored circles depict expressed (filled) and non-expressed (blank) homoeologs. Grey backgrounds illustrate cluster assignment, whereupon common surrounding depict grouping in the same co-expression cluster. **a,** Observed overall frequency distribution of homoeolog gene copies, which retain gene expression, are partially silenced (one or two homoeologs expressed) or completely absent during endosperm development (no homoeolog expressed). **b,** Total number of homoeologous triplets observed for individual clustering szenarios including Cluster I to XII and Cluster 0. **c,** Fraction of triplets are shown of which all homoeologs were placed into co-expression clusters with endosperm-specific expression patterns (Clusters I to XII only). Observed distributions were compared against the assumption of complete independence between homoeologous genes.

A total of 4,180 triplets (70%) had one or more homoeolog(s) assigned to Cluster 0, which represented unspecific gene expression during endosperm development (Fig. 5.12c). Notably, different distributions and degrees of Cluster 0-involvements were observed for triplets of which all homoeologs were found to be expressed and triplets of which one homoeolog was not detected in the grain transcriptome. For more than 80% of the triplets with three expressed homoeologous genes, at least one copy was placed in Cluster 0. On the contrary, in average two third of partially silenced triplets were associated to endosperm-specific clusters only (Cluster I to VII).

Spatiotemporal relationships in homoeologous gene expression divergence

As shown in Fig. 5.12, the majority of homoeologous genes forming triplets were placed into different co-expression clusters. Such a partitioning of triplet genes indicated differences in the gene expression profiles among homoeologous genes and clearly deviated from a naïve assumption of identical transcriptional regulation and activity for A-, B- and D-genome encoded homoeologs. To investigate for significant differences in the cluster assignments, the number of homoeologous expression transitions, i.e. different cluster assignments for homoeologous genes, was tallied in pairwise comparisons of wheat genomes and the determined co-expression clusters (Clusters I to VII) (Tables A.3 to A.5). Subsequently, the aggregation of observed transitions from one cluster to another was tested for significance by using an one-sided Fisher's exact test and Bonferroni corrected *P* values[4] (Fig. 5.13 and Table A.6).

Interestingly, clusters that are spatiotemporally related often shared a significant number of homoeologs from the same triplets. For example, a significant number of transitions of homoeologs was identified connecting early differentiation (Cluster I) and intermediate development (Cluster II) ($P \leq 0.004$) or aleurone cells sampled at 20 DPA and 30 DPA ($P \leq 0.021$). On the contrary, homoeologous transitions were only rarely observed for functionally different co-expression clusters like, for example, Cluster III (starchy endosperm) and Cluster IV (aleurone cells).

5.5 Endosperm cell type function and module-associated genome dominance

So far, this thesis found spatiotemporal partitioning of gene expression for homoeologous genes in the wheat grain transcriptome, i.e. different transcriptional activity of homoeologs in different cell types and different developmental stages during endosperm development. As the previous analyses only considered the correlation of the spatiotemporal gene expression profiles, absolute differences in transcript abundances have been disregarded. To further investigate genome asymmetry and genome dominance in terms of expression strength, the gene expression levels were directly compared among homoeologous genes.

[4] I gratefully acknowledge Karl Kugler for performing the significance test on basis of the transition matrices.

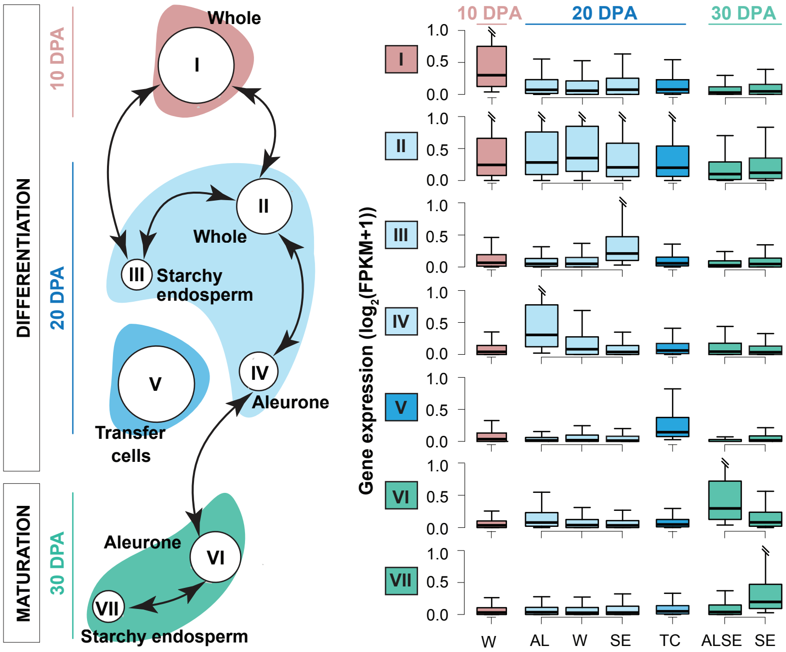


Fig. 5.13. Spatiotemporal distribution of homoeolog expression transitions.
The network arranges co-expression clusters with endosperm-specific expression profiles (Cluster I to VII represented as nodes) accordingly to developmental stages. Bi-directional arrows connecting two nodes indicate expression partitioning for a significant number of homoeologous triplets (*P* <0.05), i.e. that a significant number of homoeologous triplets have homoeolog genes located in the two connected co-expression clusters. For example, homoeologous A and B genome encoded genes were located in Cluster I, while the D genome copy clustered in Cluster II. Boxplots show gene expression profiles of the individual co-expression clusters.

5.5.1 Global patterns of homoeologous gene expression regulation

Triplet expression vectors were created by concatenating the $\log_2$(FPKM+1)-transformed gene expression values observed for the A, B and D genes forming single-copy homoeologous triplets. To determine similarities in gene expression across the spatiotemporal endosperm conditions and genomes, these vectors were combined in a matrix, which was subjected to hierarchical cluster analysis and principal component analysis. The hierarchical clustering was performed by using the `pvclust`-function implemented in the pvclust package *(318)* with Pearson's correlation distance, average linkage method and 1,000 bootstrap iterations. The principal component analysis by using the `prcomp`-command in R (parameter: scale=TRUE).

Rather then clustering of the corresponding spatiotemporal samples, the columns of the homoeologous gene expression matrix (i.e. particular endosperm cell types and developmental stages) clustered according to genomes (Fig. 5.14a). This observation was supported by strong two-dimensional separation of the genomes in the first and second principal components of the homoeologous gene expression matrix (Fig. 5.14b). Such clustering patterns indicated that, on a global scale, genome-specific gene expression dominated over cell type-specific gene expression. Notably, each genome-group maintained the spatiotemporal separation of endosperm cell types as observed in the genome-wide cluster analysis (Fig. 5.8).

No evidence was found for genome-wide transcriptional dominance of one genome. In all pairwise comparisons among the A, B and D genomes, the overall gene expression $\log_2$ fold-changes were balanced (Fig. 5.14c). However, hierarchical clustering of the triplets (correlation distance and average linkage method) partitioned the expression matrix into three segments with preferential gene expression for genes of one genomes (Fig. 5.14d). This observation indicated group-wise genome dominance and will be further investigated in more detail in the following section. Furthermore, a total of 738 triplets were differentially expressed between two genomes (A>B, A>D, B>D or vice versa) as determined by using an one-sided significance permutation test with 1,000 iterations and P values $\leq$0.05. Thereby, again, no genome-wide bias towards preferential expression of one genome was detected as a similar number of homoeologs were dominated by the A genome (232 vs. B genome and 219 vs. D genome), by the B genome (223 vs. A genome and 232 vs. D genome) and by the D genome (232 vs. A genome and 231 vs. B genome), respectively.

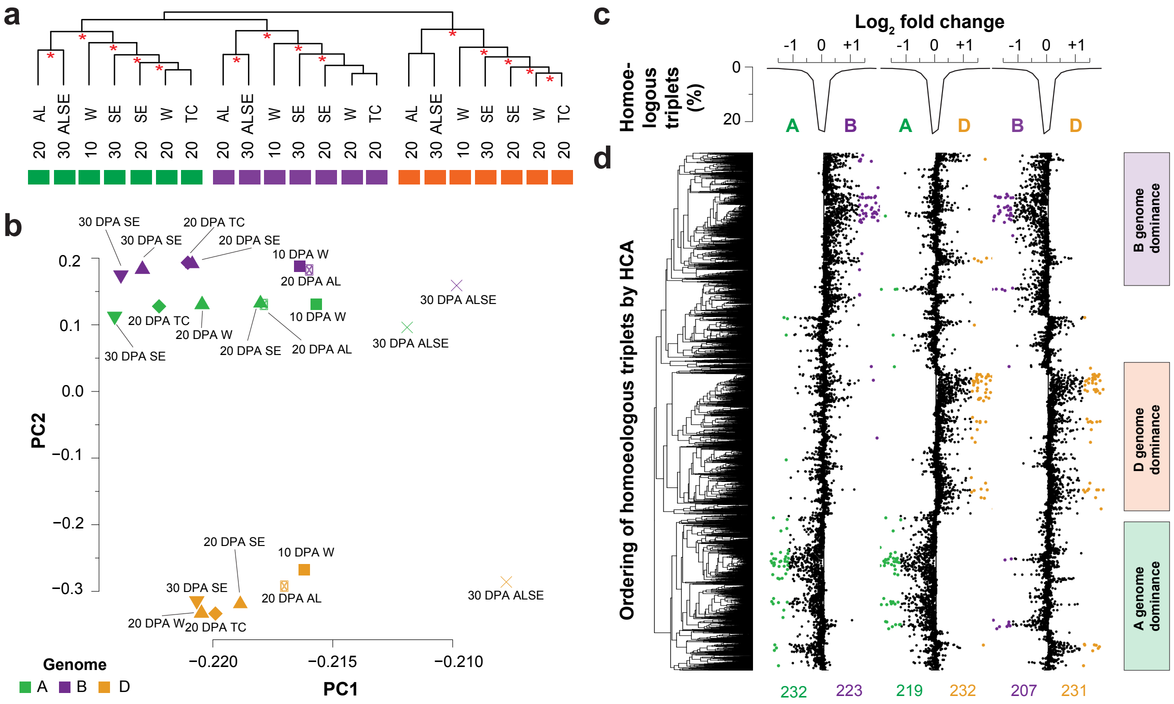


Fig. 5.14. Analysis of gene expression for single-copy homoeologous gene triplets.
Similarity in gene expression profiles was analysed for homoeologous genes of the A, B and D genes forming single-copy triplets. **a,** Hierarchical clustering across spatiotemporal conditions and genomes. Red stars mark branches with bootstrapping values above 0.9. **b,** First and second principal component identified for the homoeologous gene expression profiles. Pairwise genome comparison of mean $\log_2$ fold-changes **c,** in a genome wide analysis and **d,** for individual triplets ordered by hierarchical clustering of the gene expression matrix. Colored dots indicate significantly differentially expressed triplets (P $\leq$0.05). Right hand side boxes visualize the partitions of the dendrogram with biased genome expression.

5.5.2 Cell type and stage specific homoeologous gene expression bias

Despite absence of global deviations in the $\log_2$ fold-changes between homoeologous genes, the observed patterns in the hierarchical clustering analysis (Fig. 5.14) suggested group-wise differences in the gene expression levels for homoeologous genes. To gain deeper insights into the systems-level transcriptional dynamics of homoeologs, a further network-based co-expression analysis was conducted.

Network construction and identification of co-expression triplets

A weighted correlation network *(379)* was constructed based on the homoeologous gene expression matrix. Therefore, first, the network topology was analysed for selecting an appropriate soft thresholding power (β) to which co-expression is raised. This parameter is used for calculating the adjacency of triplets (i.e. nodes) in the network *(380)*. Different candidate power values were tested and a β of 12 selected as lowest power that reaches a scale-free topology index of 0.90 (Fig. 5.15a). Secondly, groups of closely connected genes, so called ”co-expression modules”, were identified by clustering genes based on the topological overlap matrix *(381)* and cutting the dendrogram with the `cutreeDynamic`-method *(382)* (parameters: deepSplit=2, pamRespectsDendro=FALSE, minModuleSize=50). Genes without module associations were collected in an artificial ”grey” module *(379)*. Initial modules with very similar module profiles were merged (eigengene correlation $\geq$0.75) (Fig. 5.15b). For visualization the weighted network was exported with an adjacency threshold of 0.1 to Cytoscape *(371)* (version 3.0.2) and nodes were arranged by using the ”edge-weighted force directed layout” algorithm.

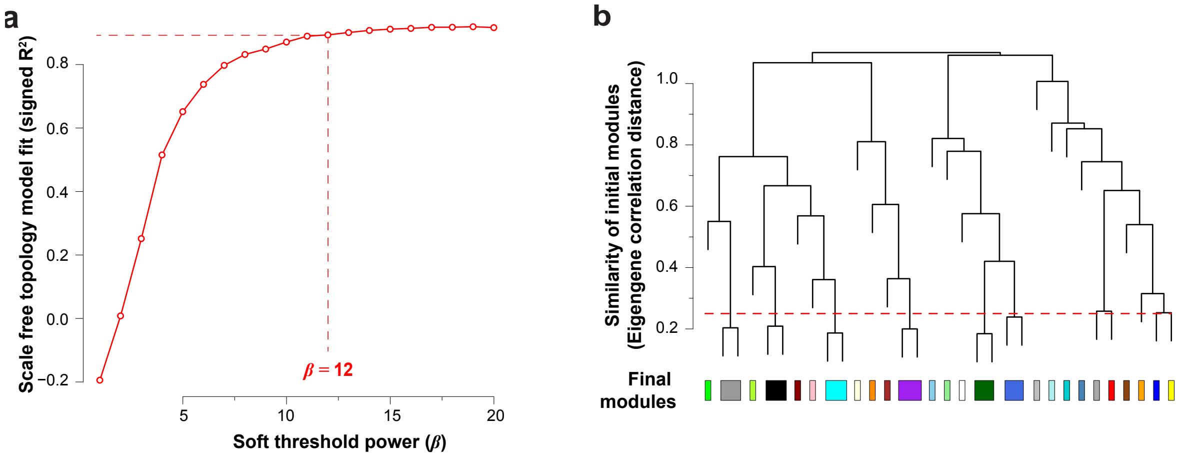


Fig. 5.15. Weighted gene co-expression network analysis for homoeologous gene triplets. Gene expression of homoeologous gene triplets was investigated utilizing network-based co-expression analysis. **a,** To identify most suitable clustering parameters different soft-threshold powers were tested and the lowest power value with a scale-free topology fit index of 0.9 was used for further analysis. **b,** Groups of genes with highly correlated gene expression, so called modules, were identified based on the co-expression network. Initial modules with highly similar gene expression profiles were merged.

Cell type specificity of the co-expression network

The computed co-expression network was partitioned into 25 clearly separated co-expression modules [Figs. 5.16 (network in inset) and A.3]. To further analyse the network topology, cell type- and stage-specificity was assigned to each network module. Therefore, the module eigengenes were correlated with pre-defined profiles specifying preferentially expression at different developmental stages and in different cell types by using the `corPvalueStudent`-function implemented in the R package WGCNA *(379)* (Fig. A.4). The assignment was then based on positive and significant correlations and integrated with information on spatiotemporal gene expression (Fig. A.5) as well as GO enrichment tests for over-representation of the molecular functions and the biological processes associated with the triplets of a module (Table A.8).

This analysis revealed module-wise expression patterns and spatiotemporal clustering in the network separating cell types and grain developmental stages (Fig. 5.16, left network). Modules related to aleurone cells (turquoise nodes) formed a large cluster of genes that were expressed at 20 DPA and 30 DPA. These were enriched for molecular functions including energy

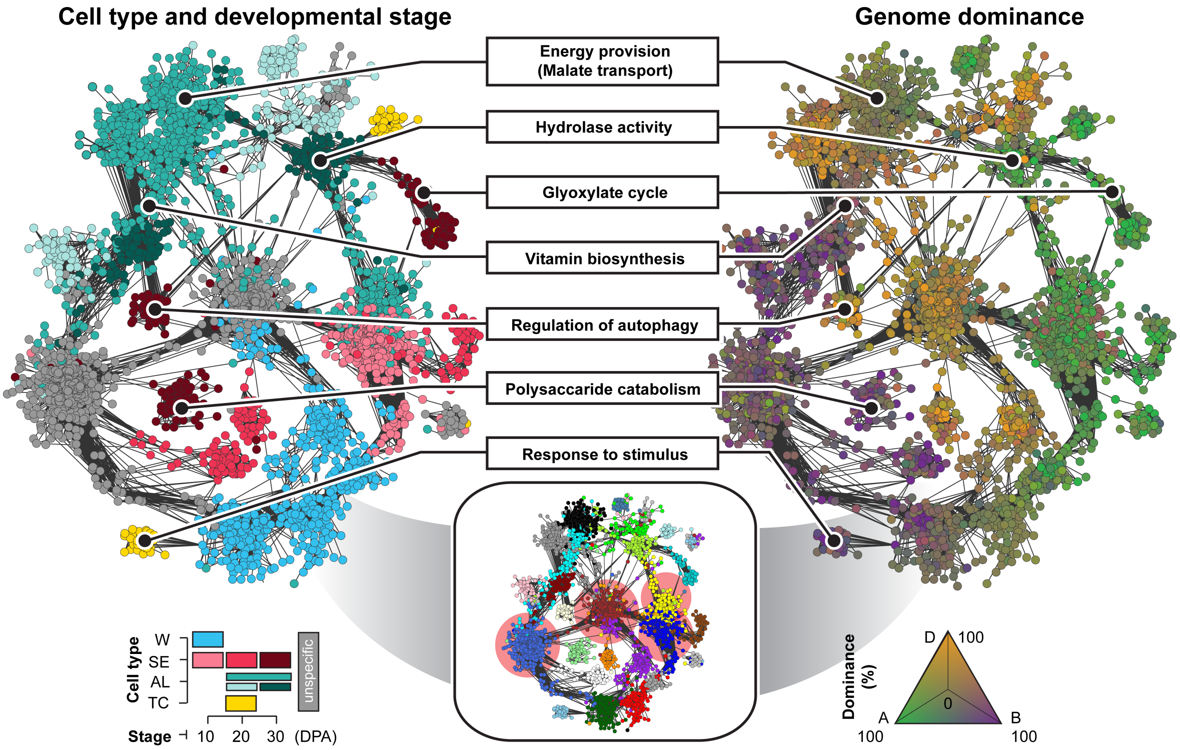


Fig. 5.16. Cell type- and developmental stage-specific gene expression and genome dominance in the homoeologous co-expression network.
25 co-expression network modules were derived by weighted gene co-expression network analysis for single-copy homoeologous gene triplets (represented as nodes). Coloring in the central inset encode for the individual network modules. Modules enriched for hub genes are highlighted by a red backgrounds. Network modules in the left panel were colored for cell type and and developmental stage based on their gene expression profiles. The coloring of nodes in the right panel visualizes genome dominance for individual triplets.

metabolism, vitamin biosynthesis and hydrolase activity. Starchy endosperm related modules (red nodes) were more scattered in the network and could be linked to polysaccharide catabolism, the glyoxylate cycle and autophagy. Transfer cells (yellow nodes) formed dense, separated clusters enriched for ”response to stimulus” functionality. Transcriptional modules enriched for more general functionalities (e.g. transport and translation) without cell type or developmental phase specificity were also found (grey nodes). In total modules with aleurone-specific expression patterns constituted more than one third of the nodes (2,207), whereas the other cell types contributed to a lesser extent [658 nodes for starchy endosperm (11%) and 149 for transfer cells (2%)]. The remaining nodes grouped either with the early phase of endosperm development or unspecific clusters.

Genome dominance for co-expression modules

Besides comparison of gene expression levels, the genome dominance of each module was assessed by using an enrichment test for significantly high numbers of DE homoeologous genes applying a one-sided Fisher's exact test (Bonferroni corrected P value <0.05) (Fig. A.3). Furthermore, to visualize genome dominance in the correlation network, individual nodes were coloured by a weighted mean of the genome-specific average expression across all samples (Fig. 5.16, right network). Different genomes dominated expression for 23 of the modules, which included 92% of the analysed homoeologous triplets. Notably, no single genome proved to be overly dominant corroborating the observations made with the hierarchical cluster analysis above.

Generally, highly connected nodes, so called ”hubs”, display characteristic expression profiles for network modules *(383)*. Therefore, the hub nodes of the inferred co-expression network were defined as those triplets within the top 10$^{\text{th}}$ percentile of a centrality measure computed with the igraph package *(384)*. Noteworthy, significantly more hub genes showed a homoeologous gene expression bias than non-hub genes did (one sided Fisher's exact test with Bonferroni corrected P value <0.05). Moreover, each module was assessed for enrichment of hub genes (one sided Fisher's exact test with Bonferroni corrected P value <0.01). This revealed significant enrichment for modules that served as connecting layers among different regions of the network [Figs. 5.16 (highlighted modules in the inset) and A.3]. Considering the special roles of hub genes in co-expression networks *(383)*, these observations suggested that the identified hubs might play an important role in orchestrating genome-specific expression in the grain co-expression network.

Functional compartmentalization of the bread wheat transcriptome

Furthermore, the observed genome asymmetry and cell type specificity were superimposed with a semantic aggregation of significantly over-represented GO categories (Table A.8). Therefore, a two-dimensional semantic distribution was computed for all biological process GO categories that were over-represented in any transcriptional group by using the REVIGO webserver *(363)*. Subsequently, the distribution was compared among groups with asymmetric genome expression by coloring those terms green, purple or orange, which were significantly over-represented in transcriptional groups dominated by the A, B or D genome, respectively.

This visualization strategy revealed functional specialization and compartmentalization with subdivision of basic cellular functions as well as endosperm-specific functions for individual cell types or developmental stages among the wheat genomes (Fig. 5.17). Fundamental functions related to translation and DNA repair were dominated by the A genome, whilst B genome-dominated groups were enriched in genes related to chromosome organization and D genome-dominated groups were enriched for transport activity or signal transduction. Favored expression of one genome for specific spatiotemporal endosperm conditions was also found. For instance, lipid metabolism was dominated by the A genome, monosaccharide metabolism dominated by the B genome or catabolic processes and authophagy dominated by the D genome.

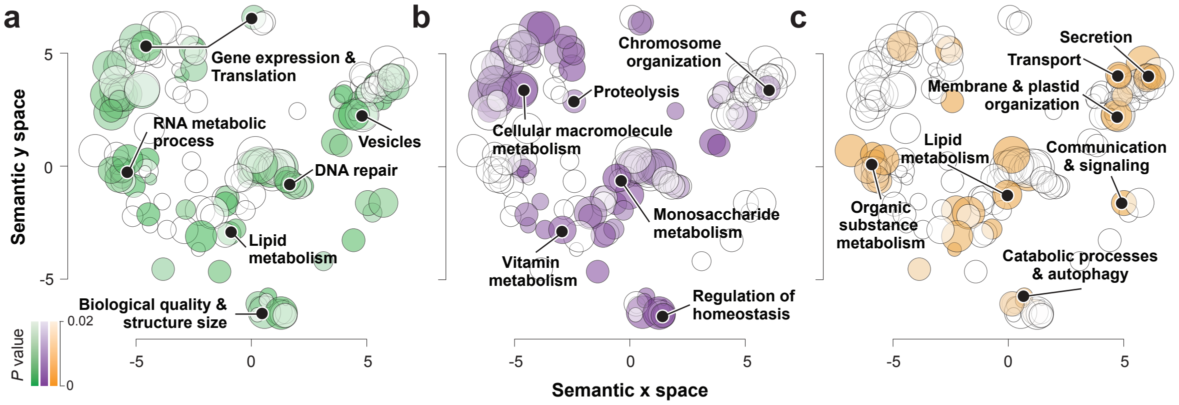


Fig. 5.17. Functional compartmentalization of homoeologous gene expression.
Semantic similarities between significant enriched GO terms (biological processes) of all triplet co-expression groups were calculated and projected onto a two-dimensional semantion space using *(363)*. GO categories were colored, if they have been identified to be significantly over-represented (*P* <0.02) in any of the subgenome-dominanted triplet co-expression groups. **a,** A genome, **b,** B genome and **c,** D genome.

5.5.3 Sequence evolution vs. expression evolution

In Section 4.3.4 of this dissertation, evidence of incongruence in the gene family composition was observed on chromosome arm level, which corroborated recent genome-wide studies that investigated the phylogenetic relationships within the Triticeae and suggested non-linear, reticulated evolution of the A-, B- and D-genome lineages *(69,71,73)*. To elucidate if the evolutionary history of homoeologous genes relate to genome asymmetry in gene expression, the sequence-based features of homoeologous genes were compared with transcription-based features. Therefore, sequence divergence analysis was conducted on basis of the number of synonymous substitutions per synonymous site (K_s) and the number nonsynonymous substitutions per nonsynonymous site (K_a). These measures are proxies for the divergence in protein sequences (K_a), the evolutionary relationships and distances (K_s) and selection pressure (K_a/K_s). For each triplet the best scoring protein BLAST *(227)* ($E < 10^{-10}$) alignment was determined between homoeologous wheat genes and the K_a, K_s, and K_a/K_s values were computed *(281,282)*.

The levels of synonymous substitutions per synonymous site between A-D and B-D homoeologous gene pairs were comparable and significantly smaller than for pairs from the A and B genomes [Wilcoxon-Mann-Whitney-Test (*P* <0.001)] (Fig. 5.18). This pattern corroborated observations of Marcussen *et al.*, who reported variation in phylogenetic relateness of the A, B and D genomes and higher frequency of *B(A,D)* and *A(B,D)* topologies *(69)*. However, same overall evolutionary patterns were found for individual co-expression modules and independent from gene expression level dominance (Figs. 5.18 and A.6).

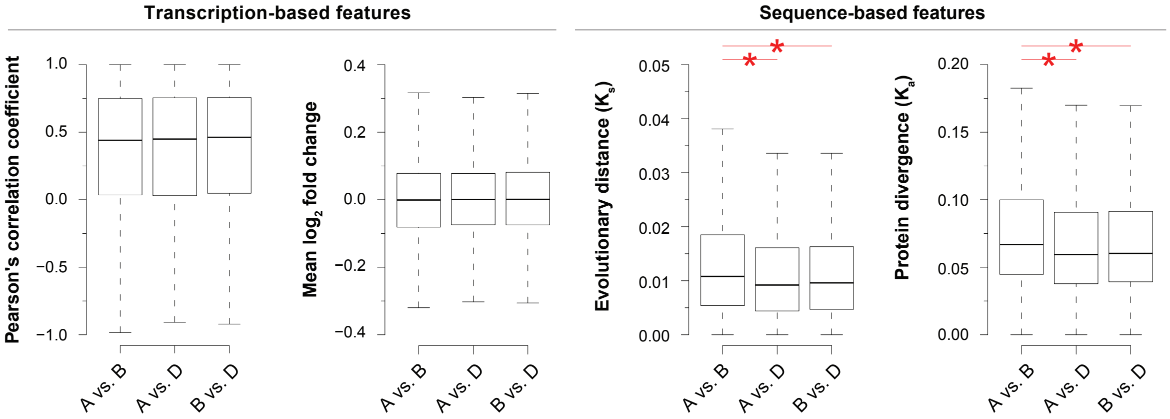


Fig. 5.18. Comparison of transcriptional and sequence-based differences for homoeologous genes. Transcriptional and sequence-based features were compared between genome-pairs for all homoeologous gene triplets. Red stars mark significant differences in distributions of sequence-based features [Wilcoxon-Mann-Whitney-Test (*P* <0.001)]. Corresponding pairwise comparisons separately for triplets dominated by the A, B or D genome are shown in Fig. A.6.

5.6 Chromosomal regulation of wheat gene expression

In the previous sections, this work revealed significant expression differences between homoeologous genes causing transcriptional genome asymmetry that related to non-random subdivision of functional responsibilities. Such homoeolog-specific expression patterns may have been set already in the di- or tetraploid progenitor genomes and were inherited by hexaploid wheat or, alternatively, were set following genome merger *(101,133)*. Recent single gene studies suggested that those changes might be a result of genetic and epigenetic regulatory mechanisms, which orchestrate gene expression in stochastic as well as non-stochastically modes *(108,385)*. However, while genetic regulatory mechanisms are expected to affect genes at different genomic locations, epigenetic mechanisms often influence neighbouring genes *(106,386,387)*. The impact of chromosomal position on gene expression will be further investigated in this section.

5.6.1 Construction of Triticeae prototype chromosomes

In absence of a fully sequenced and ordered reference genome for bread wheat, first, the annotated wheat genes were projected into a sequential ordering. Therefore, the ”crop circle” model *(34,35,55–58)*, which describes large conservation of synteny and gene order in the grasses (Section 1.2.3), provided a powerful principle and allowed the comparative genomics-based construction of seven ”Triticeae prototype” (Tp) chromosomes. The anchoring of wheat genes along the Tp scaffolds reflected the virtual ancestral linear gene order of the A, B and D genomes. This approximation allowed the comparative analysis of positional gene expression regulation between homoeologous chromosomes.

Construction of the Triticeae prototype scaffolds

Barley and bread wheat diverged approximately 13 mya (Section 1.2.2). As large conservation in genome structure exists between these two species *(388)*, the chromosomal ordering of more than 21,000 barley genes *(62)* served as a suitable proxy for the definition of syntenic regions between wheat and the high-quality reference grass genome sequences of *Brachypodium (42)*, rice *(45)* and sorghum *(40)*. To construct the Triticeae prototype scaffolds, I extracted the genes contained in syntenic chromosomal segments from each reference genome. These segments were linearly placed correspondingly to the ordering in the barley genome *(62)*. Within each segment, the reference genes were arranged based on the closest evolutionary distance to wheat, i.e. *Brachypodium* genes were ordered first and rice and sorghum genes were successively added (Fig. 5.19a, steps i to v). Thereby, putative orthologous genes were assigned to a common prototype locus as determined in pairwise best bidirectional protein BLAST *(227)* searches between the *Brachypodium*, rice and sorghum gene sets (steps ii and iv) ($E < 10^{-5}$). The remaining genes, which miss an orthologous counterpart, were arranged next to an anchored gene with minimal genomic distance (steps iii and v).

A total of 21,956 *Brachypodium*, 22,916 rice and 20,738 sorghum genes were identified to be located in syntenic regions that are unambiguously designated to one barley chromosome (Table 5.5). These genes were integrated into 37,608 loci along seven Tp chromosome scaffolds ranging between 4,133 (chromosome 6) up to 6,169 (chromosome 2) loci. Overall, 11,349 putative orthologous relationships were determined between all three reference species and 5,304 orthologs pairs were found between two species only. These were anchored to same loci in the prototype backbones (Fig. 5.19b). The extracted blocks with syntenic conservation were clearly apparent from structural comparison between the Tp chromosomes and the *Brachypodium* genome (Fig. 5.19c).

Anchoring of wheat genes at Triticeae prototype scaffolds

To anchor bread wheat genes along the Tp chromosome scaffolds, the predicted HC proteins were aligned against the entire gene sets of *Brachypodium*, rice and sorghum, respectively (BLASTP *(227)* with E value $\leq 10^{-5}$). Considering only the best-scoring alignment with mini-

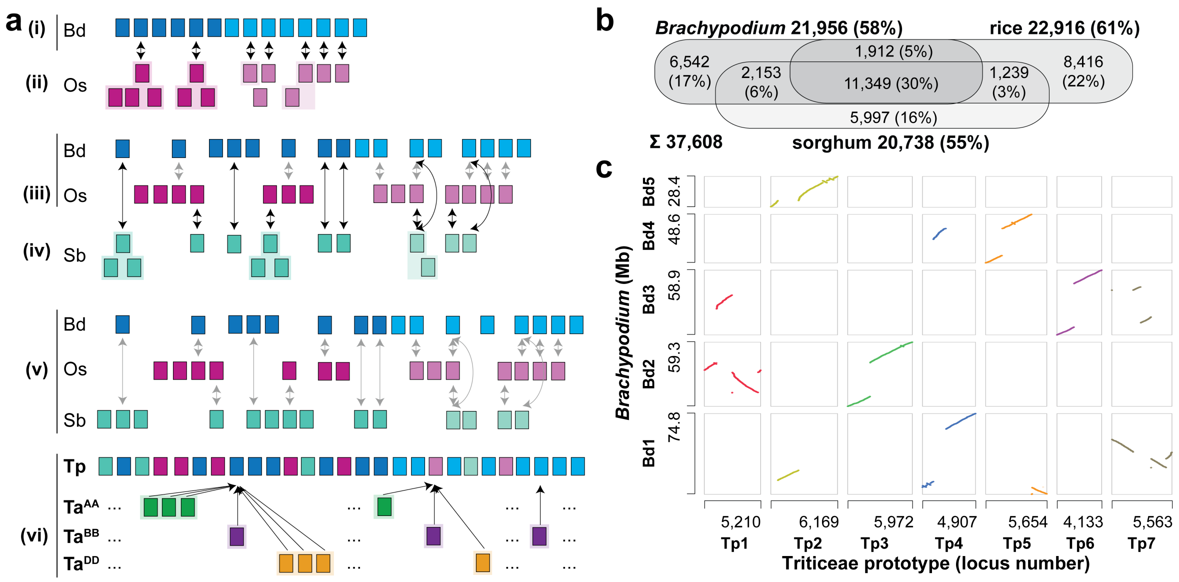


Fig. 5.19. Construction of the Triticeae prototype chromosomes.
a, Workflow for construction of seven Triticeae prototype (Tp) chromosomes. For each chromosome genes located in syntenic regions were extracted from the *Brachypodium* (Bd, blue), rice (Os, red) and sorghum (Sb, turquoise) genome based on comparisons to the barley gene order (the color shading of rectangles indicates corresponding blocks). See main text for description of individual steps (i)-(vi). **b,** Number of defined orthologs between *Brachypodium*, rice and sorghum (overlaps in Venn diagram) and singletons that were integrated in the Tp scaffolds. **c,** Comparison of the seven Tp chromosomes to the *Brachypodium* genome. Maximum locus number and physical position is shown for each Tp and *Brachypdodium* chromosome, respectively.

Table 5.5. Number of *Brachypodium*, rice and sorghum genes building the seven Triticeae prototype chromosome scaffolds.

Species	Tp1	Tp2	Tp3	Tp4	Tp5	Tp6	Tp7	Σ
Brachypodium	3,108	3,647	3,391	3,003	3,135	2,538	3,134	21,956
rice	3,158	3,806	4,003	3,306	2,991	2,651	3,001	22,916
sorghum	2,746	3,276	3,471	2,242	3,217	2,473	3,313	20,738
Σ loci	5,210	6,169	5,972	4,907	5,654	4,133	5,563	37,608

mum 65% alignment identity spanning at least 30 amino acids length, the wheat genes were associated to a reference gene in the Tp scaffolds accordingly to nearest evolutionary distance (Section 1.2.2).

Overall, more than two third of HC bread wheat genes [57,903 genes (HC1 to HC3)] were positioned along the seven ancestral chromosomes with a similar proportion of genes from the A genome [18,778 genes (70%)], the B genome [20,479 genes (67%)] and the D genome [18,646 genes (67%)] (Fig. 5.20a). While a comparable number of genes of individual genomes were assigned to chromosomes 1 and 4, the anchoring of wheat genes sightly differed for the other chromosomes. Consistent with the extraordinary size of chromosome 3B *(241,311)* the largest differences among homoeologous chromosomes were observed for this group. A combined set incorporating 16,286 ordered Tp loci was defined (Fig. 5.20b), of which 59% (9,629) were supported by genes from all three wheat genomes and 22% (3,506) by combination of two wheat genomes. For all pairwise combinations similar overlaps between the A, B and D genomes were observed suggesting no predominant deletion or retention of genes from one genome.

By using this strategy most wheat genes were assigned to their corresponding chromosome in the Tp (Fig. 5.20c). However, some structural re-arrangements were observed, which are not present in the barley genome on which basis the Tp was built (Fig. 5.20c, highlighted regions). For example, these included a translocation between chromosome 4 and 5, which is shared by all homoeologous chromosomes, and the two well-described translocations between chromosomes 4AL/5AL and 7BS/4AL *(63,315)* (Section 4.3.4). Furthermore, a previously unknown deletion in the short arm of chromosome 6D was present that will be further discuss in the following section of this chapter (Section 5.7). Excluding these local re-arrangements, the number of anchored genes along the Tp chromosomes did not largely deviate for individual genomes. Small local regions with an extraordinary number of anchored genes were caused by *Brachypodium*, rice and sorghum genes that are classified as transposable elements and, thus, led to an increased number of anchored bread wheat genes attracted from all chromosomes.

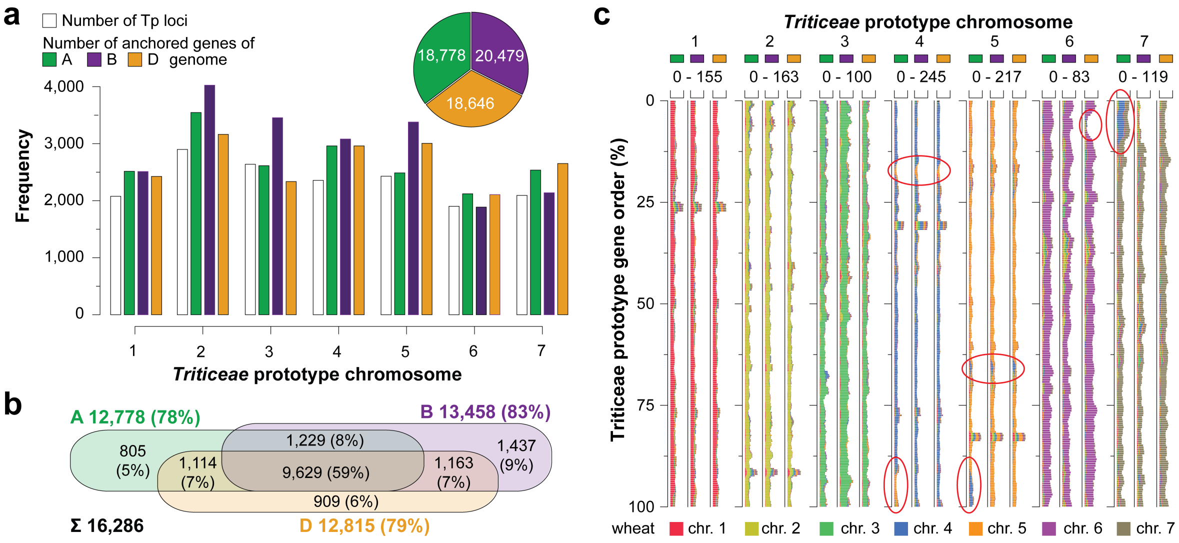


Fig. 5.20. Anchoring statistics of wheat genes to the seven Triticeae prototype chromosomes. **a,** Number of anchored wheat genes to the seven Tp chromosomes. Pie chart visualizes the total number of anchored genes. **b,** Number of Tp loci that are supported by genes *Brachypodium*, rice or sorghum. The intersections visualize the number of identified orthologs between one, two or all three analysed species. **c,** Number of bread wheat genes anchored per window along each Tp chromosome (sliding window including 50 loci, window shift size of 10 Tp loci). Color code of bars indicate the chromosome of the anchored genes in the wheat genome. Minimum and maximum number of integrated genes per window is shown for each Tp chromosome. Regions highlighted in red represent chromosomal re-arrangements in wheat that were not shared with the barley genome.

The Tp chromosomes represented a simplified approximation of the present linear gene order in the bread wheat genome, which was also apparent from comparisons to the wheat GenomeZipper generated in frame of the IWGSC consortia *(70)* (Section 4.3.2). Generally, a high structural agreement was found between the Tp and the GenomeZipper for each chromosome (Fig. 5.21). The Tp considered only chromosomal re-arrangements that were common to the A-, B- and D-genome lineages, whereas discarded genome-specific and small-scale interruptions in microsynteny due to neglecting genetic marker information. Although this underestimated structural variation among the three wheat genomes, however, this simplification allowed one-to-one comparative analyses between corresponding homoeologous chromosomes. By using the wheat GenomeZipper the most likely location in the wheat genome can be inferred for further analysis.

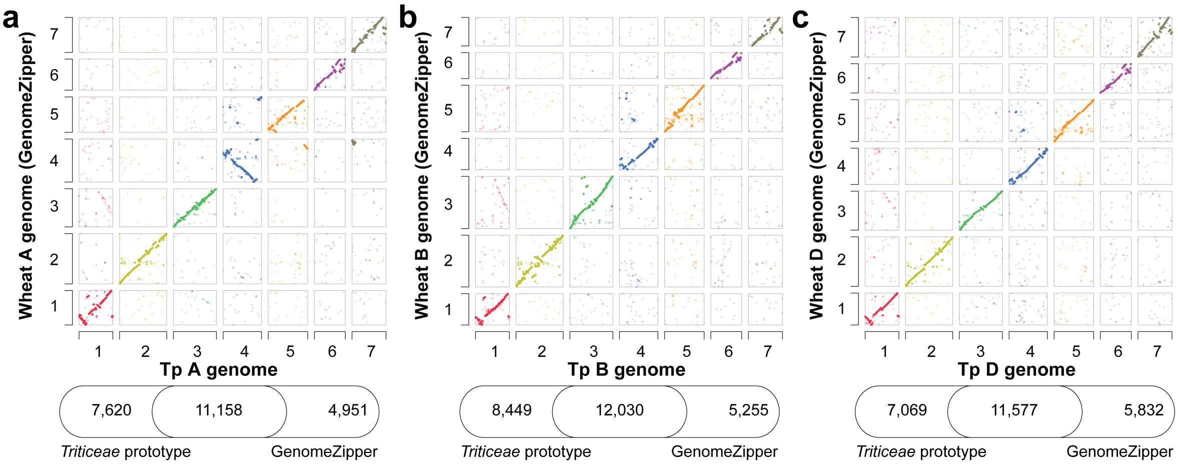


Fig. 5.21. Structural comparison between the Triticeae prototype against the wheat GenomeZipper. For each genome dotplots visualize the position of bread wheat genes (HC1 to HC3) in the seven Triticeae prototype chromosomes and in the wheat GenomeZipper for the **a,** A genome, **b,** B genome and **c,** D genome. Venn diagrams count the number of wheat genes that were anchored by one or both approaches.

5.6.2 Chromosomal regulation of endosperm gene expression

To elucidate the chromosomal effects on mRNA abundances, gene expression was measured along the Titiceae prototype chromosomes by using a sliding window approach (median expression strength for windows including 50 Tp loci and 10 Tp loci window shift size). Along all chromosomes gene expression oscillated and chromosomal domains were found with increased transcriptional activity during wheat grain development (Figs. 5.22 and A.7 to A.13). Generally, the spatial patterns of chromosomal gene expression differed only minor between endosperm cell types and developmental phases. Furthermore, the observed patterns were to a large extent similar between genomes. However, various chromosomal segments showed apparent divergent expression patterns between endosperm cell types and developmental stages as well as between wheat genomes. As exemplified for two domains in the following, such differences might

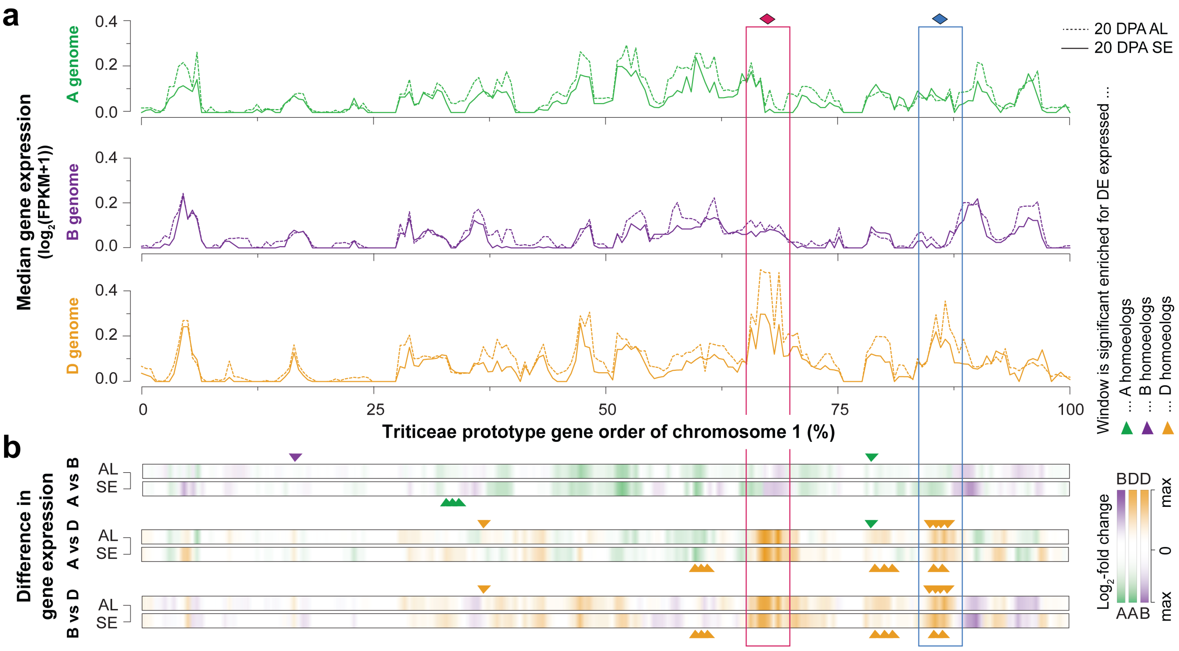


Fig. 5.22. Chromosomal regulation of gene expression along the Triticeae prototype gene order exemplified for chromosomes 1.
a, Local regulatory divergence between homoeologous gene exemplified by Triticeae prototype (Tp) chromosome 1 (sliding window, size 50 Tp loci; shift 10 Tp loci). Line charts show the median gene expression measured in aleurone and starchy endosperm cells at 20 DPA. **b,** Pairwise $\log_2$-fold changes in gene expression for each window between wheat subgenomes. Triangles indicate chromosomal regions that are significantly enriched for homoeologous triplets up-regulated in a single genome (Fisher's exact test with *P* value <0.05).

be triggered and influenced by numerous factors *(108,385)*. Therefore, domains with asymmetric expression patterns constitute potential targets for elucidating the underlying silencing and enhancing mechanism in full detail and further studies.

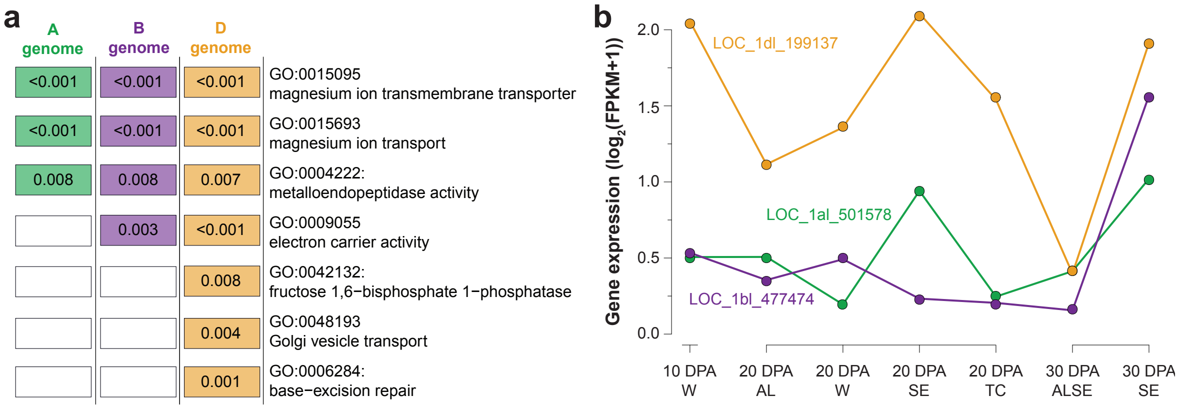
Local deviations in gene content

One of these domains, which is indicated by a red diamond in Fig. 5.22, was located on the long arm of Triticeae prototype chromosome 1. Gene expression significantly differed across endosperm samples with highest abundance in aleurone cells [Wilcoxon-Mann-Whitney-Test (*P* <0.01)]. D-genome encoded genes dominated expression over genes of the A and B genomes, which were similarly expressed. A total of 169 expressed wheat genes were anchored within this chromosomal region, of which 27 genes were significant differentially expressed between samples (FDR <0.05). Correspondingly to the expression profile most of these DE gene were exclusively up-regulated in aleurone cells [6 genes of the A genome, 4 genes of the B genome and 7 genes of the D genome (17 genes)]. These encoded proteins involved in major processes and pathways of wheat endosperm including gluconeogenesis, lipid binding and gibberellin signal transduction, an important hormone acting in development and growth control *(389)*. Slightly more genes of the D genome (61 genes) were located within this chromosomal segment compared to the A (56 genes) or B (52 genes) genomes. Considering the significantly enriched functional categories for these genes, various GO terms were exclusively found over-represented for the D-genome encoded genes, which indicated that the corresponding proteins were only present in this genome (Fig. 5.23a). Interestingly, these genes function in processes and pathways characteristic for aleurone cells including gluconeogenesis (GO category ”Fructose-1,6-bisphosphatase”) *(390)* or vesicle-mediated transport *(391)*. Therefore, the in-balance and the local variation in gene content might have caused the observed asymmetric gene expression profile between the A, B and D genomes for this chromosomal domain.

Regulatory mechanisms acting on particular regions of homoeologous chromosomes

As a second example, the chromosomal domain indicated by a blue diamond in Fig. 5.22 showed also considerably increased gene expression in the D genome and similar expression in the A and B genomes. Differentially expressed homoeologous triplets, which were dominated by the D genome, were found to be significantly over-represented in this segment (one sided Fisher's exact test with Bonferroni corrected *P* value <0.05). This excluded that variations in the local gene content between genomes caused the observed gene expression differences. Moreover, it indicated that common regulatory mechanism controlled the transcriptional activity for homoeologous genes present as a single-copy in each genome.

For instance, one triplet with asymmetric gene expression was related to DnaJ chaperone proteins. All homoeologous followed a generally similar expression pattern and showed maximum expression in SE-containing samples (Fig. 5.23b). This finding agreed with the known function and expression patterns for this class of proteins, which was associated with important roles in en-



A genome	B genome	D genome	
<0.001	<0.001	<0.001	GO:0015095 magnesium ion transmembrane transporter
<0.001	<0.001	<0.001	GO:0015693 magnesium ion transport
0.008	0.008	0.007	GO:0004222: metalloendopeptidase activity
	0.003	<0.001	GO:0009055 electron carrier activity
		0.008	GO:0042132: fructose 1,6−bisphosphate 1−phosphatase
		0.004	GO:0048193 Golgi vesicle transport
		0.001	GO:0006284: base−excision repair

Fig. 5.23. Exemplified analysis of chromosomal domains with non-balanced gene expression. **a,** Functional enrichment analysis for gene ontology categories encoded by wheat genes positioned in the red-marked chromosomal domain in Fig. 5.22 (position 66-69%). Significant over-represented GO terms are shown for individual wheat genomes (*P* <0.01). **b,** Expression level across spatiotemporal endosperm samples for homoeologous genes related to DnaJ class of chaperone proteins. Genes are located in the blue-marked chromosomal domain in Fig. 5.22 (position 84-87%).

dosperm and protein body development and was found to be highly abundant in the sub-aleurone, where starch and storage protein accumulation occurs *(392)*. However, correspondingly to the observed patterns in this chromosomal domain, the D-genome encoded homoeologous gene copy significantly dominated gene expression compared to the transcriptional activity of the A- and B-genome encoded copies.

5.7 Targeted expression profiling of gene families affecting wheat baking quality

White flour is the major contributor to humankind's nutrition and one of the main ingredients for most wheat products. Cereal seed proteins have not only nutritional importance, moreover, they facilitate the biotechnological process of breadmaking *(393)*. During milling seeds are separated into their individual components. Wheat bran (aleurone cells, seed coats and pericarp) and the embryo are removed, whereas the starchy endosperm is disposed for further flour manufacturing. Starchy endosperm is the largest body of wheat grains and accumulates the majority of seed and starch proteins mostly in form of prolamins, which attribute the texture and unique characteristics of wheat dough permitting to bake bread *(353,394,395)*. Prolamins are derived from glutamine and rich in proline and amide nitrogen and, in wheat, can be divided into two functionally different components, the gliadin and glutenin proteins. By adding water to flour these two proteins forms a complex called gluten, which built-up a network structure that is simultaneously extensible and elastic (”visco-elasticity”). Under action of baking powder or yeast gluten permits dough to stretch and rise and, thus, is responsible for the texture of bread *(393)*.

The wheat prolamin gene family shows enormous genetic diversity, differs largely between wheat cultivars and includes hundreds of genes and extensive allelic variation, which composition controls and influences bread baking quality of individual wheat varieties *(394)*. Therefore, a global understanding of the constitution of gene families affecting dough quality is required for improving wheat varieties. In this study, the bread wheat genome was screened for genes affecting baking quality, namely the prolamins, which include the high molecular weight glutenin (HMW) and low molecular weight (LMW) glutenin (Glu) genes, the α-, γ-, and ω-gliadin (Gli) genes *(353,394)*, the grain hardness (Ha) locus, which includes the puroindoline A (*pinA*) and puroindoline B (*pinB*) genes *(396,397)*, and the storage protein activator (SPA) proteins *(398)*.

5.7.1 Cataloguing genes affecting wheat baking quality

Public cDNA and protein sequence information deposited in the NCBI sequence database[5] was utilized to target candidate genes in the bread wheat gene annotation as well as in the CSS

[5] HMW-Glu: *(43,399)*; LMW-Glu: *(43,400)*; *pinA*, *pinB* and *pinB2*: *(43,401)*; SPA: *(402)*; α-Gli: O. D. Anderson, direct submission to NCBI (U50984.1), *(403)*; ω-Gli: *(404)*; γ-Gli: *(344,405)*

assembly by using manual BLAST *(227)* and GenomeThreader *(305)* searches. Furthermore, OrthoMCL *(223)* was applied to cluster proteins of the bread wheat, the *A. thaliana (137)* and *Ae. tauschii (43)* genomes into gene families [OrthoMCL *(223)* (version 2.0) using BLASTP ($E \leq 10^{-5}$) and an inflation parameter of 1.5]. The individual evidences were manually combined and annotated bread wheat genes associated to the corresponding target gene families. The glutenin and gliadine genes have highly complex protein-coding sequences including multiple repetitive protein domains *(406)*, which substantially complicated a full-length assembly of the gene loci by using NGS-based genomic resources. Therefore, structures and sequences of some genes and transcripts were curated by hand with respect to the alignments of query sequences and wheat RNA-seq transcriptome information.

For each individual gene family the orthologous relationships were investigated by using multiple protein sequence alignments [CLUSTALW algorithm *(407)*] and phylogenetic trees [”neighbourhood joining” algorithm with ”average percent identity” method implemented in the Jalview software *(408)* (version 2.8)]. Due to the complex sequence composition *(406)*, some of the analysed grain quality genes have not been assigned to the high-confidence gene classes utilizing the IWGSC gene annotation pipeline. As initially only the HC gene set was considered for calculation of expression levels in this study (Section 5.2.5), the mRNA abundances were re-computed for all genes of the analysed grain quality gene families. Therefore, the number of RNA-seq reads falling within the curated gene structures were counted by using HTSeq *(409)* and, subsequently, gene expression strength calculated in RPKM (Reads Per Kilobase exon model per Million mapped reads) by normalizing the read counts to the total number of mapped reads for individual replicates and the transcript length *(180)*. The final gene expression levels of the analysed genes were defined as mean RPKM across all biological replicates.

5.7.2 Gene family compositions and gene expression patterns for seed and storage proteins during endosperm development

The high- and low molecular weight glutenin genes

Glutenins are polymeric proteins *(410)* and mainly contribute to dough characteristics *(411)* like elasticity and strength through the formation of disulfide bonds *(412)*. Based on differences in their molecular mass glutenins were classified into the high molecular weight glutenin subunit and the low molecular weight glutenin subunit *(394)*, which presence, allelic variations and expression strength have been associated with dough properties and superior bread-baking performance *(342,411,413)*. The LMW-Glu subunit genes were found on the short arm and HMW-Glu subunit genes on the long arm of the group one homoeologous chromosomes (Figs. 5.24a and b, dendrograms). Six genes were identified for the HMW-Glu subunit occurring as homoeologous triplets, which could be further classified into two homoeologous groups representing the x-type (1×A, 1×B and 1×D genome) and the y-type (1×A, 1×B and 1×D genome) HMW glutenins. The two homoeologous groups differed in phylogenetic topology showing closer similarity between the B and D homoeologs for the x-type and between the A and B homoeologs for the y-type. On the contrary, twelve genes of the LMW-Glu subunit showed non-balanced occurrence across genomes (2×A , 5×B and 5×D genome). Nine of these genes had intact open reading frames (1×A, 3× B and 5× D genome), while the protein sequences of three LMW-Glu genes were interrupted by premature stop codons due to frame shifts or nonsense mutations (1×A and 2× B genome). Generally, these findings were in good agreement with previous characterizations of the HMW and LMW gene families in the bread wheat cultivar "Chinese Spring" utilizing PCR-based techniques or proteomics *(400,414,415)*. However, one intact LMW-Glu gene reported in the D genome was not found *(400)*, but instead a novel glutenin gene (*GluB3-**) was identified in the B genome with closest sequence homology to the *Glu-B3* gene.

Overall, the HMW-Glu and LMW-Glu genes were most abundant in starchy endosperm (Figs. 5.24a and b, heat maps), which was consistent with the accumulation of starch and storage

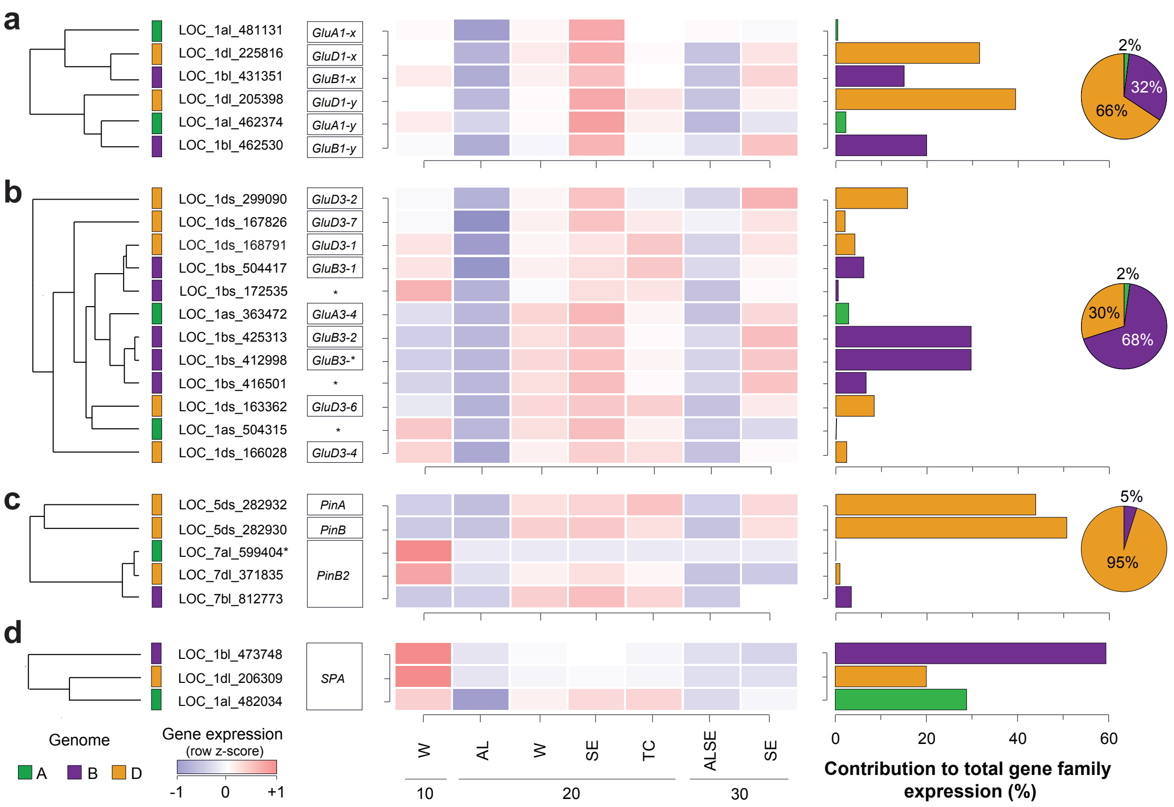


Fig. 5.24. Analysis of members of the glutenin and puroindole gene families and the seed storage protein activator genes.
Major seed storage proteins contributing to baking quality of bread wheat were identified for **a,** the high molecular weight glutenin subunit (HMW-Glu), **b,** the low molecular weight glutenin subunit (LMW-Glu), **c,** the puroindoline *PinA*, *PinB* and *PinB2* genes and **d,** the storage protein activator (SPA) genes. Dendrograms depict a phylogenetic tree for each gene family. Relative gene expression levels (row z-score) are visualized as heat maps. Bar and pie charts visualize the relative contribution of individual genes and genomes to total gene family expression. Stars label putative pseudogenes with interrupted by premature stop codons, frame shifts or repetitive elements.

proteins in these cells. However, LMW-Glu genes showed also increased mRNA abundances in the transfer cell sample (20 DPA TC). This sample included tightly attached surrounding starchy endosperm cells, which could not be completely removed in the dissection process (Fig. 5.1). Therefore, the SE-contamination most likely caused the observed expression patterns, largely line with findings of Tosi *et al. (343)*, who reported a gluten protein gradient in the wheat endosperm with higher abundance of LMW-Glu genes in the close sub-aleurone region. Besides similar spatial expression patterns, temporal differences in gene expression of certain subunits were also present. Whilst all six HMW-Glu subunit genes were predominantly expressed at intermediate endosperm development (20 DPA), transcription LMW-Glu genes changed over time. In contrast to *GluB3-1*, *GluD3-1* and *GluD3-4*, for which decreasing mRNA levels were measured over developmental time, *GluA3-4*, *GluB3-2*, *GluB3-**, *GluD3-2*, *GluD3-6* and *GluD3-7* showed increased gene expression at intermediate and late endosperm development. Thereby, similar spatiotemporal gene expression profiles were, at least in parts, reflected in the phylogenetic relationship of LMW-Glu genes across genomes (e.g. *GluB3-1* and *GluD3-1*).

For both glutenin gene families individual wheat genomes contributed differently to overall gene family expression (Fig. 5.24a and b, bar and circle diagrams). Whereas the total gene expression of the LMW-Glu was dominated by the B genome (68%), genes of the D genome accounted for two third of total expression of the HMW-Glu subunit. Generally, A genome-encoded genes contributed only marginally (2%), which supported the observed inactivation of *Glu-A* locus in hexaploid wheat *(416)*. The three identified putative pseudogenes were transcribed, although at a considerably reduced level compared to intact proteins from the B and D genome, respectively.

The puroindoline grain hardness locus

An additional major contributor to baking quality is the Ha locus controlling the physical characteristics of the endosperm texture, which differentiate cultivated pasta wheat (*T. turgidum*, hard endosperm) from hexaploid bread wheat (soft and hard endosperm varieties) *(417)*. The Ha locus was exclusively detected on the short arm of chromosome 5D encoding for the puroindole A and puroindole B genes *(418)* (Fig. 5.24c). The presence of the puroindolines on only one wheat genome was consistent with the evolutionary fate of this locus and constituted the D genome a special contribution to the kernel structure of bread wheat grains *(419)*. The *PinA* and *PinB* genes have been reported for the diploid A, B and D genome progenitors, but were absent in tetraploid wheat *T. turgidum* and, thus, also in the A and B genomes of bread wheat. By the hybridization of tetraploid wheat and the D genome progenitor *Ae. tauschii*, the Ha locus was integrated back into bread wheat genome *(341,420)*. In addition, a second locus, *PinB2*, was identified on the short arm of chromosomes 7A, 7B and 7D, respectively. This locus showed approximately 70% protein sequence homology to *PinA* and *PinB* proteins *(401)*. *PinB2* genes were also associated with differences in kernel texture and wheat yield traits *(421,422)*. However, besides substantial differences in protein sequences, the transcriptional activity of *PinB2* largely differed from the puroindoline locus on 5DS. In agreement with experimental results *(421)*, *PinA* and *PinB* genes were expressed at high levels at intermediate and late endosperm development (20 DPA and 30 DPA), whereas the *PinB2* homoeoalleles were expressed substantially less and predominantly at early endosperm development (10 DPA).

The storage protein activator proteins

Storage protein activator (SPA) genes play a crucial role in orchestrating expression of grain storage proteins in wheat and have been correlated with grain hardness *(398)*. Three SPA gene copies were present, one in each genome on the long arm of chromosome 1 (Fig. 5.24d). The gene copies of the B and D genomes were predominantly expressed only at 10 DPA, whereupon the A genome derived SPA allele at 10 DPA and 20 DPA (W, TC and SE). This behaviour was consistent with Wan *et al. (423)*, who reported temporal variations in the gene expression for SPA homoeologs and also decreasing transcriptional activity over time. On the contrary to the counterbalanced presence of homoeologous SPA genes, asymmetric gene expression resulted in dominance of the B genome over the A and D genomes. However, in contrast to the HMW-Glu, LMW-Glu and puroindolines, the A genome contributed substantially to SPA gene family expression (24%).

The gliadine gene family

The gliadins account for up to 40% of total wheat flour and, thus, are important contributors to human diet *(415)*. As they are key initiators of celiac disease, an autoimmune disorder, understanding sequence composition and expression of these genes is of industrial and also medical importance *(344,415,424)*. As shown in Fig. 5.25a, this study revealed substantial variations in the number of genes as well as in the relative expression levels between the A, B and D genomes. Many of the query proteins could only be partially aligned against the CSS assembly. Some of these gene fragments showed considerable deteriorated protein sequences and contained in-frame stop codons, which indicated highly dynamic gene family composition including a substantial degree of pseudogenization.

γ-Gli and ω-Gli gene candidates were identified on the short arm of group 1 chromosomes. B and D genome-encoded gene copies dominated total gene family expression, while copies of the A genome were less abundant in the wheat grain transcriptome. The α-Gli genes were encoded on the short arm of chromosome 6 and, strikingly, only candidate genes were found in the A and B genomes. No α-Gli query could be aligned to D genome-contigs in the underlying genomic reference assembly of the braed wheat genome. Most likely, this originated in a previously undescribed deletion of approximately 200 genes in the short arm of chromosome 6D (Figs. 5.25b and 5.20). As the α-gliadine locus has been identified in *Ae. tauschii*, the diploid progenitor of the D genome *(43)*, the findings suggested a ("Chinese Spring") bread wheat-specific deletion of this chromosomal segment. Interestingly, in contrast to the γ- and ω-gliadins as well as both glutenine subunits, the A genome contributed essentially to the total expression of α-gliadins.

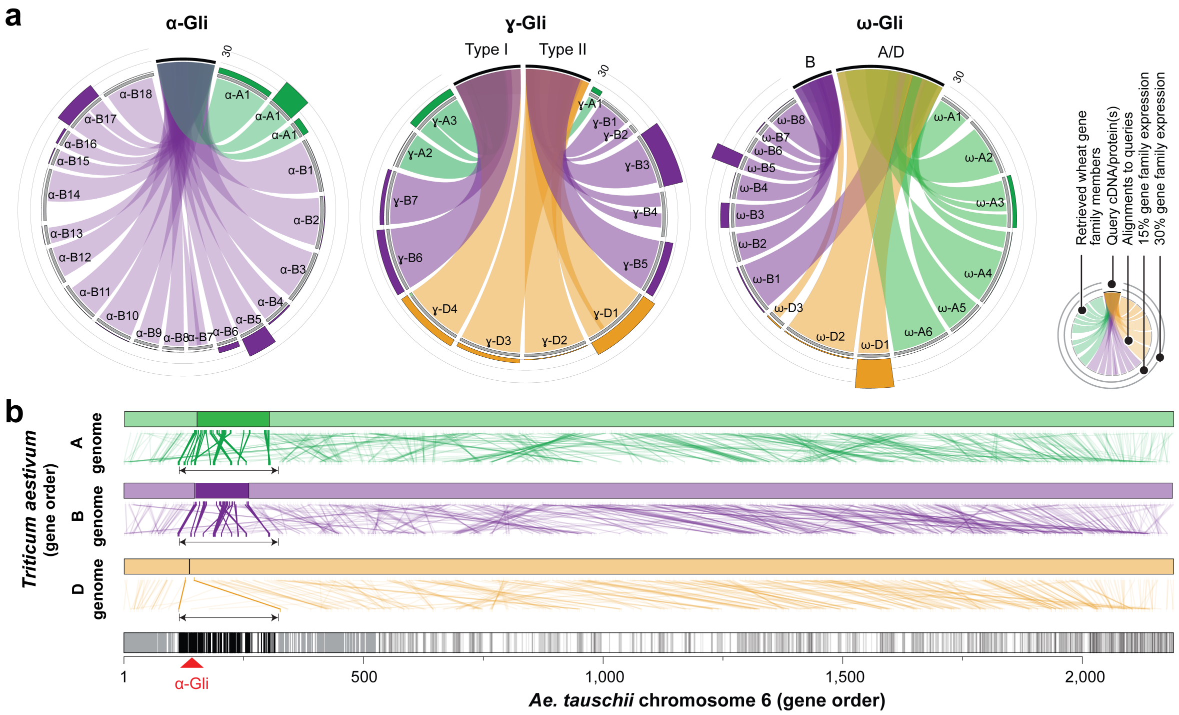


Fig. 5.25. Analysis of the α-, γ- and ω-gliadin gene families.
a, Identified gene(-fragments) for the α-, γ- and ω-gliadin gene families in the CSS assembly. Query sequences are indicated by black bars and detected gliadins by grey bars. Coverage of query genes is indicated by the connectors. Bar heights in the outer circle visualize the relative contribution of all samples to overall gene family expression. **b,** Structural comparison between *Ae. tauschii* and each bread wheat genome. Links indicate location of putative orthologous gene pairs between *Ae. tauschii* (black bar) and bread wheat (colored bars). Bold connectors highlight the deleted segment on chromosome 6DS.

5.8 Discussion

High-throughput mRNA sequencing technology allowed to monitor gene expression in one of the most important organs, the bread wheat grain. The nuclear endosperm was separated into the major endosperm cell types at three developmental stages spanning the differentiation to maturation phases. On basis of the IWGSC CSS draft genome sequence assembly and gene annotation bioinformatic analysis of more than 1.6 billion paired-end mRNA-sequencing reads (>160 Gb) was conducted to investigate the wheat endosperm transcriptome on multiple levels. Starting from a global profiling of the transcriptional activity in individual cell types and time points, spatiotemporal co-expression clusters were identified and the regulatory mechanisms and functional aspects of homoeologous gene expression divergence and genome asymmetry were elucidated with a network-based approach for an important polyploid cereal.

5.8.1 Highly complex and flexible alternative splicing in bread wheat

Complementary to the structural gene annotation of the CSS assembly (Chapter 4), this work aimed at studying the expression of genes and splicing variants during endosperm development. In addition to the existent gene annotation the underlying data set provided evidence for 401 previously undefined high-confidence gene loci (Table 5.1). The majority of these genes fell into the HC4 gene set and represented most likely gene fragments or pseudogenes. Only five genes were classified as functional, full-length gene predictions and were added to the HC1 class. These low numbers indicated saturation in the general characterization of (protein-coding) gene loci and corroborated the high completeness of the bread wheat gene catalogue annotated in frame of the IWGSC project (Chapter 4).

However, the detection of only a few previously undefined genes contrasted with abundance and annotation of more than 15,000 novel alternative splicing forms mainly located at existing IWGSC wheat gene loci (Table 5.1). Genes with novel splicing variants encoded for a broad range of gene ontology categories incorporating basic cellular processes as well as endosperm-specific functions (Fig. 5.4). The majority of transcripts was preferentially expressed in individual cell types or at particular time points (Fig. 5.7), which suggested a highly complex and flexible splicing landscape in wheat. This corroborated recent findings in plants *(184,191)* and mammals *(180,183,425)* attributing fundamental roles and high impact to splicing regulators in the control of cellular protein composition. Therefore, alternative splicing constitute to a considerably increase in protein diversity and may provide a reservoir of different gene products for a broad range of functions and pathways *(307,323)*. In accordance with the argumentation of Reddy *et al. (308)*, the large number of previously unknown transcript predictions highlighted the importance of deep transcriptome profiling to identify tissue-, cell type-, time point- or environmental-specific splicing variants, which constitute potential targets for an in-detail investigation of expression regulation by alternative splicing.

5.8.2 Large differences in spatiotemporal gene expression patterns of wheat endosperm

Across the tested spatiotemporal samples approximately half of the high-confidence wheat genes were found expressed during wheat grain development (Fig. 5.7), which was largely consistent with previous observations in *Arabidopsis (356)* and barley *(358)*. Preferentially expressed genes represented only a minor fraction of the entire wheat transcriptome with a maximum of 644 for aleurone cells at the intermediate developmental stage (20 DPA AL) (Table 5.4). A low number of PEGs has been previously observed in *Arabidopsis (356)*, thus, the results suggested conserved regulatory principles in grain development across more than 100 million years of plant evolution.

The functional and morphological differences between aleurone cells, starchy endosperm and transfer cells were also evident from the gene expression measurement (Fig. 5.8). Aleurone cells grouped apart from the samples including starchy endosperm, which accorded well with early initiation of cell type specification before 10 DPA *(346)*. Starchy endosperm-containing samples grouped and higher cluster distances were observed among SE samples reflecting different developmental phases. Moreover, the majority of the identified significant differentially expressed genes were exclusively up-regulated for individual time points or for subsequent time points (Fig. 5.9). This corroborated functional progression and specification in endosperm development towards grain maturation over time *(346,356,426)*. A large number of exclusively up-regulated genes in 10 DPA indicated pronounced transcriptional activity and dynamics in early endosperm development, while considerably less genes were up-regulated in 30 DPA SE, consistent with the initiated cell death and decrease in gene expression in mature starchy endosperm cells compared to aleurone that actively participate in nutrition of the growing embryo *(354,355,370)*.

Endosperm development progresses through four phases: the syncytial and cellularization phases (<10 DPA), the differentiation phase (10 to 20 DPA) and the maturation phase (>30 DPA) *(216,217)*. Co-expressed genes were grouped into seven clusters characterizing gene expression in the two latter phases, in which the industrially important characteristics of wheat grains are set. These co-expression clusters showed specific spatiotemporal expression profiles, which were designated for individual cell types (Cluster VI: aleurone cells; Clusters III and VII: starchy endosperm; and Cluster V: transfer cells) and developmental stages (Cluster I: early endosperm development; Cluster III, IV and V: intermediate endosperm development; and Clusters VI and VII: endosperm maturation) (Fig. 5.11). Furthermore, the functional annotation of genes grouped in each co-expression cluster revealed enrichment for biological processes, which were consistent with literature and fit to the observed gene expression profiles for each cluster. Thereby, aleurone cells and starchy endosperm functionally shifted and reprogrammed gene expression over time. At 20 DPA aleurone cells mainly expressed proteins that encoded for lipid and carbohydrate metabolism, but proteins involved in transmembrane transport were predominant at 30 DPA. This agreed well with initiation of aleurone cells to produce and release enzymes for mobilizing nutrients to the germinating embryo in the maturation phase *(346,354,355)*. Consistent with the expected initiation of programmed cell death of SE cells at around 30 DPA *(370)*, the majority of expressed genes in starchy endosperm encoded for proteins involved in autolysis and related apoptosis pathways.

5.8.3 No global transcriptional dominance for wheat genomes

Previously, absence of a suitable reference genome sequence and high similarity of the homoeologous gene copies impeded distinguishing contribution of individual wheat genomes to entire transcriptome composition *(346)*. In this study, the combination of the CSS assembly and gene annotation with next generation RNA-sequencing enabled to profile homoeolog-specific expression patterns and provided the opportunity to investigate the organization of gene expression in hexaploid wheat. On a global scale, a similar number of expressed genes was found in the A, B and D genome across all sampled cell types and time points (Fig. 5.7). This was consistent with the balanced genome structure and gene content described in the previous chapter of this thesis (Chapter 4). No bias towards one or the other wheat genome was evident in terms of preferentially expressed genes (Table 5.4) nor for significant differentially expressed genes (Fig. 5.9). This indicated generally conserved and balanced contributions of the A, B and D genomes to important features and functional pathways. This contrasted to transcriptional dominance of one progenitor genomes, which has been reported for duplicated genes derived from ancestral whole genome duplications in allopolyploid cotton *(116,427)* or paleotetraploid maize *(111)*.

5.8.4 Subfunctionalization of homoeologous genes

Microarray analysis of syntentic wheat allopolyploids have also demonstrated that rapid changes upon polyploidization cause differential expression of homoeologous genes and non-additive gene expression patterns *(125)*. To test for transcriptional differences between the three wheat genomes, the expression profiles were compared for 6,576 single-copy homoeologous gene triplets with exactly one gene copy from each genome. Overall, a strong conservation of gene expression was observed as all three or none of the homoeologs were expressed for the majority of the triplets (Fig. 5.12). Silencing of one or two copies occurred less often then expected by chance. However, partitioning into different co-expression clusters was found for genes forming homoeologous triplets, an observation suggesting a substantial level of divergence in the gene expression profiles between copies origin from different genomes. This largely agreed with findings of Mochida *et al. (123)* and Bottley *et al. (124)*, who reported organ-specific expression and silencing of homoeologous genes (Section 1.3.3).

A significant number of expression transitions occurred only between spatiotemporally related clusters, which represented same cell types (e.g. whole endosperm and aleurone cells at 20 DPA) or connected adjacent developmental phases (e.g. whole endosperm at 10 DPA and 20 DPA) (Fig. 5.13). The predominance of non-radical alterations in the spatiotemporal dimension indicated that expression subfunctionalization, rather than neofunctionalization, is the major evolutionary mechanism underlying expression divergence in the three bread wheat genomes. Furthermore, as for a considerable number of triplets one or two homoeologous genes were placed into the ”zero” cluster, while the other homoeolog(s) were found in the endosperm-specific co-expression modules (Fig. 5.12). This might indicate loss of spatiotemporally specific gene expression as an intermediate stage on the way to silencing or expression divergence. Vice versa, such patterns could also reflect cell type- or stage-specific divergence and subfunctionalization of homoeologous genes.

A similar numbers of gene expression transitions were found for homoeologous genes in all pairwise genome comparisons. Interestingly, this observation did not correspond to the longer common evolution of the A and B genomes in the tetraploid progenitor genome. Therefore, the observed divergence could be either pre-existing in the diploid parents of the A-, B- and D-genome lineages and maintained during common evolution in the polyploid or, alternatively, resulted from reprogramming of gene expression in the hexaploid genome. It is noteworthy that the observed degree of differential expression among homoeologs between genomes was not correlated to the time of the polyploidization events.

5.8.5 Homoeolog gene expression divergence and functional genome asymmetry

The observed divergences in the grouping of homoeologous genes to co-expression clusters considered only correlation in spatiotemporal gene expression and disregarded differences in the absolute mRNA abundances between homoeologous genes. Direct comparison of gene expression levels and hierarchical cluster analysis of homoeologous gene expression revealed striking autonomous regulation of wheat genomes, each maintaining the overall pattern of gene expression similarity during endosperm development (Fig. 5.14). Again, the conserved genome-specific patterns strongly contrasted with patterns of gene expression in older polyploids or in rediploidized genomes, where one of the genomes was found to be more transcriptionally active than others *(111,113,133,134)*.

However, while on a global scale gene expression differences seemed to be balanced between genomes, a network-based co-expression analysis identified 25 groups of homoeologous triplets, which showed substantial bias towards up- or down-regulation of individual genomes (Fig. 5.16). These co-expression modules were associated with distinct cell types and developmental stages of wheat endosperm. Different combinations of genome asymmetry were observed with a comparable amount of groups dominated by either the A, B or D genome. Furthermore, central genes were preferentially differentially expressed between genomes indicating important function in the control and the orchestration of polyploid genome expression.

Interestingly, functional enrichment analysis of co-transcribed groups revealed that at least part of the expression divergence between genomes reflected subdivision of cellular functions among wheat homoeologs (Fig. 5.17). This suggested that genome dominance is not the result of a random process. Rather it follows a concerted schema and might be related to mechanisms that function between genomes to balance expression of individual and groups of genes *(428)*. Differences related to specific functional gene categories imply that both, fundamental cellular processes as well as major features of bread wheat grain development, were attributable to contributions from single genomes.

Sequence divergence analysis of homoeologous gene pairs revealed significant increased evolutionary distance between the A and B genomes in comparison to the A and D genomes as well as the B and D genomes, which had similar distributions for the number of nonsynonymous substitiutions per synonymous site. These patterns supported the hypothesis of incongruent evolution in the Triticeae *(71,73)* (Section 4.3.4) and homoploid hybrid speciation of the D-genome lineage *(69)*. Evolutionary signals did not affect homoeologous gene expression in terms of spatiotemporal profiles, i.e. differences in expression correlation, genome asymmetry and expression level dominance (Fig. 5.18). Even further, the absence of correlation among transcriptional activity and sequence-based phylogenetic signals suggested non-sequence-related genetic or epigenetic regulatory mechanisms orchestrating gene expression among the three homoeologous wheat genomes *(108,385)*.

5.8.6 Chromosomal regulation of wheat gene expression

A high degree of regulatory orchestration between genomes, but simultaneous maintenance of autonomous genome expression patterns, has been attributed to the evolution of *cis*-regulatory elements coupled to epigenetic mechanisms *(106,387)*. The positional effects on gene expression were investigated and, in absence of a yet complete and ordered reference genome sequence, almost 60,000 wheat genes anchored along seven Triticeae prototype chromosomes (Fig. 5.20). These were built on basis of the "crop circle" model *(55)* utilizing conserved synteny to *Brachypodium*, rice and sorghum and represented the virtual, ancestral state of Triticeae genomes. Excluding known re-arrangements like, for example, the translocation of segments between chromosomes 4A, 5A and 7B *(63,315)*, and a previously unknown deletion on the short arm of chromosome 6D, high degree of structural conservation between homoeologous chromosomes and no preferential retention, gain or loss of genes from one genome were evident (Fig. 5.20). Therefore, the projections of wheat genes from individual genomes onto a common genomic axis allowed the direct comparison of positional gene expression between homoeologous chromosomes by disregarding genome-specific and small-scale structural differences.

Variations in gene expression levels along chromosomes resulted in chromosomal domains preferentially expressed during wheat grain development *(429,430)* (Fig. 5.22). Although the chromosomal distribution of gene expression was largely synchronized between endosperm cell types and developmental stages as well as between the A, B and D genomes, various chromosomal domains showed asynchronous expression patterns. These domains constitute potential targets for further specific analysis of the underlying biological causative mechanisms of genome asymmetry. As exemplified for two of these domains, this study indicated that a complex mixture of genetic and epigenetic factors may regulate the expression of homoeologous genes *(118,120)* (Fig. 5.23). Genetic differences led to locally diverged gene expression between the A, B and D genomes. Variations in genome compositions resulted in a non-balanced set of encoded biological functions, which may were caused by lineage- or genome-specific gain or loss of genes and implied asymmetric contribution of individual genomes to the entire transcriptome. In addition, genome-dominated chromosomal domains accumulated a significant number of differentially expressed homoeologous gene triplets. The presence of single-copy homoeologs in each genome precluded expression differences to be caused by local gene copy number variations. It demonstrated that there is local genome asymmetry for neighboring genes common to the A, B and D wheat genomes. This allowed to speculate that genome level dominance, at least in parts, might be caused by epigenetic regulatory mechanisms, which may act differently on particular corresponding domains of homoeologous chromosomes.

5.8.7 Dominance of the B and D genomes for genes affecting baking quality

Genes contributing to the unique visco-elastic characteristics of bread wheat are of special agricultural, industrial and medical importance, since presence and expression of certain subunits were associated to superior bread-making performance *(342,413,424)*. Here, genes of the prolamin gene families (glutenin and gliadine genes), the puroindole genes as well as the seed storage activator proteins were analysed. So far, absence of a (draft) reference genome sequence has impeded such a genome-wide investigation for bread wheat. Previous studies were mainly focussed on single genes or gene-families utilizing, for example, PCR-based techniques, proteomics or long-read sequencing technology *(344,400,403,414,415,431)*. This work provided a comprehensive reference gene catalogue of genes affecting baking quality combining sequences, structural annotation and transcriptional activity to aid breeding of high quality bread wheat varieties.

Consistent with single-gene analyses, large differences were found in sizes and constitutions for the investigated gene families. Substantial variations in the phylogenetic relationships and non-balanced contribution of the A, B and D genomes were present, often including deteriorated gene fragments and putative pseudogenes (Figs. 5.24 and 5.25). Starchy endosperm cells showed highest transcriptional activity of baking quality genes. Thereby, the total gene family expression was mostly dominated by genes of the B and D genomes, whereas the A genome contributed only marginally. Pseudogene candidates showed evidence for expression, but at a substantially lower level than intact gene copies, which indicated down-regulation of disrupted proteins along with functional deterioration.

This study also revealed a small deletion on the short arm of chromosome 6D (Fig. 5.25), which has not been reported for the analysed bread wheat cultivar previously. Interestingly, this deletion included the α-gliadin gene locus and, contrary to all other gene families, the α-Gli genes encoded in the A genome substantially contributed to total family expression. Regarding this transcriptional behaviour, which was atypical for the analysed gene families, the particular contribution of the A genes might be triggered by the loss of the D genome counterparts following a pattern observed in syntentic tetraploid wheat, as discussed by Feldman *et al. (135)*.

5.9 Conclusions

This study represented a major improvement for bread wheat towards a genome-wide understanding of gene expression in different organs, cell types, developmental stages or in plants grown under or exposed to various different environmental conditions. Different aspects of endosperm development was investigated on a genome-wide level and preferentially expressed genes and co-expression modules were identified. The findings revealed a complex interplay in gene expression regulation during grain development in hexaploid bread wheat that involves several layers. Globally, genes of the A, B and D genomes contributed similarly to the wheat endosperm transcriptome. Genome-specific expression was found to dominate over tissue-specific expression, which suggested a considerable degree of autonomous regulation of the homoeologous wheat genomes. However, substantial divergence in expression profiles of homoeologous genes indicated spatiotemporal subfunctionalization, asymmetric contribution of genomes to particular functions as well as different regulation in particular domains of homoeologous chromosomes.

The wide and unpredictable variation in wheat quality and yield caused by both genetic and environmental factors as well as their interaction, represent severe challenges to the wheat industry. The resources and techniques developed in this thesis form an important basis for addressing the inter- and intragenomic regulation within a polyploid genome. This study provided a reference gene catalogue that enables studying the functional output in a wide range of wheat cultivars and under different environmental regimes to allow the identification of the underlying genetic and epigenetic factors and their interplay in wheat. This will impact the improvement of agronomical and industrial traits of one of the world's most important crops and contribute to ensure global food security.

Chapter 6

Summary and perspectives

Recent advantages in DNA sequencing technologies have tremendously changed the field of plant genome and transcriptome analysis. Increased availability of genomic resources opened new dimensions for plant breeding and accelerated the identification of new varieties with improved yields and improved resistance to challenging environmental conditions *(432–435)*. However, sequence assembly and analysis of plant genomes still face severe difficulties. Especially, large genome sizes, highly repetitive DNA sequences and polyploidy delayed the construction of (draft) reference genome sequences for most Triticeae including allohexaploid bread wheat, one of the world's major cereals. By using heterogeneous data sets generated by NGS methods in different international consortia, this thesis focussed on the development and application of bioinformatic strategies to establish resources necessary to overcome current limitations in wheat genome research. The implemented approaches deepened the knowledge about structure, constitution and organization of the allohexaploid bread wheat genome and enabled genome-wide investigation of the evolutionary fate of homoeologous genes and the impact of polyploidization on spatio-temporal expression patterns in the developing endosperm. The scientific achievements made throughout this thesis contributed novel insights for a global understanding of the complex genome interplay with genetic and likely epigenetic regulatory mechanisms orchestrating gene expression of an important, polyploid cereal.

Establishment of comprehensive genome resources

Two complementary approaches were used, which allowed establishing a genome-wide overview of genes with an assignment on chromosome and genome level. This work highlights the importance of comparative genomics-based bioinformatics strategies exploiting orthologous gene family relationships among grass genomes and, thus, facilitating stringent homoeolog-specific assembly of whole genome shotgun sequences that covered essentially the entire gene repertoire of hexaploid wheat. In addition, the implementation of an extrinsic gene prediction, suitable for the annotation of NGS assemblies and complex plant genomes, allowed defining the coding sequences and the transcript structures for more than 90% of bread wheat genes. All generated data resources have been made publicly available and will support the isolation of agronomically important genes and gene families for further systematic research supporting breeding strategies for improved wheat varieties.

Limited, unbiased gene loss in hexaploid bread wheat

Based on independent methods utilizing different data resources the bread wheat genome was estimated to contain between 94,000 to 106,000 protein-coding genes. Comparative gene family analysis between allohexaploid wheat and the diploid D-genome progenitor *Ae. tauschii* as well as related grass genomes indicated pronounced retention of homoeologous gene, especially for single-copy gene families, and overall limited gene loss following polyploidization. On a global level, the predicted genes were similarly distributed among the A, B and D genomes and not preferentially retained or deleted in one particular genome. However, differences were evident between homoeologous chromosomes including sporadic structural rearrangements, small-scale chromosomal deletions as well as incongruent patterns in the constitution of individual gene families. These observations suggested pre-existing differences in the parental genomes or, alternatively, that evolutionary mechanisms act differently on single chromosomes or individual chromosomal regions.

Genome dynamics following polyploidy

Although retention of homoeologous genes and structural conservation were observed, this thesis also revealed a dynamic evolution of the bread wheat genome following polyploidy. Various gene families with an expanded number of genes in hexaploid wheat were identified and associated with processes that function in cellular organization and control as well as with pathways underlying agricultural important traits. These genes represent candidates for further targeted analysis. Moreover, the abundance of gene fragments related to key functions in adaptive responses to environmental stimuli and abiotic stresses suggested gene duplication as an essential mechanism potentially providing a reservoir for rapid adaption to environmental changes.

Extensive alternative splicing and post-transcriptional expression regulation

Extensive alternative splicing was observed for A-, B- and D-genome encoded genes with similar frequencies of distinct splicing types in each wheat genome. Furthermore, a substantial number of premature termination codon-containing splicing variants was identified, which indicated that a considerable proportion of genes might be post-transcriptionally regulated via nonsense-mediated decay and the RUST pathway. High tissue-specific expression of splicing variants and significant conservation of PTC$^+$/NMD candidate genes among homoeologous genomes supported recent observations in other species attributing important roles to splicing-based expression regulation to increase proteome range.

Homoeologous expression bias associated with functional compartmentalization

Insights into genome-wide spatiotemporal gene expression patterns on homoeologous resolution in allohexaploid bread wheat were gained. By using deep RNA sequencing the transcriptional activity in the endosperm cell types during the differentiation and maturation phases was profiled demonstrating a high degree of regulatory autonomy for the three wheat genomes. While no global transcriptional bias towards the A, B or D genome was evident, network-based co-expression analysis for single-copy homoeologos gene triplets indicated a cell type- and stage-dependent homoeologous gene expression bias. Preferential transcript abundances in either genome was associated with distinct cellular functions and biological processes and suggested functional compartmentalization of the wheat transcriptome.

Gene expression regulation is linked to chromosomal domains

Synteny-based construction of seven Triticeae prototype chromosomes approximated the ancestral positional gene order in the wheat A, B and D genomes. Gene expression oscillated and formed chromosomal domains, which were to a large extent synchronized among cell types, time points and genomes. Presence of domains with asynchronous patterns suggested a common regulatory mechanism for co-localized genes. Thus, epigenetic modifications might differently affect particular chromosomal regions and contribute to gene expression dominance.

Genome asymmetry in gene families of agronomic and industrial importance

Targeted analysis of gene families affecting baking quality revealed genome asymmetry exists for agronomic and industrial important traits. The prolamin genes were catalogued highlighting specific patterns for individual gene families, such as large variations in gene copy number, frequent pseudogenization, genome-specific chromosomal deletions, genome-dominance and expression bias as well as dosage compensation effects in absence of homoeologous counterparts.

Concluding remarks and perspectives

This dissertation aimed at making use of next generation sequencing to work towards reference genome resources for complex plant genomes. The observed patterns suggested a complex interplay of genetic and epigenetic mechanisms, potential trans-regulatory mechanisms and cross-talk between genomes impacting the allohexaploid bread wheat genome. However, the origin, causative principles and involved regulatory pathways still need to be determined and require additional experiments, specific data sets and further bioinformatic-driven analysis. Such studies will rely on data sets generated within this work, which enable a comprehensive genetic and functional analysis, provide starting points to analyse the regulatory principles controlling the fate of homoeologous genes and allow a systematical study of the homoeolog's contribution to traits of agricultural or industrial importance. Additional comprehensive genome, transcriptome and methylation sequencing for multiple tissues of bread wheat and related di- and tetraploid wheat genomes or synthetic crosses between species would allow to distinguish sporadic and repeatable alterations, to determine short-, mid- and long-term consequences following polyploidzation, to distinguishing between genetic and epigenetic changes as well as to identify the key compounds orchestrating inter- and intra-genomic expression in a polyploid genome.

Bibliography

[1] D. J. Spielman, R. Pandya-Lorch. Proven successes in agricultural development. A technical compendium to Millions Fed. *International Food Policy Research Institute, Washington, D.C.*, 2010.

[2] A. Evans. The feeding of the nine billion: global food security for the 21st century. *Royal Institute of International Affairs*, 2009.

[3] D. K. Ray, N. D. Mueller, P. C. West, J. A. Foley. Yield trends are insufficient to double global crop production by 2050. *PloS One*, 8(6):e66428, 2013.

[4] FAO, IFAD, WFP. The state of food insecurity in the world 2013. The multiple dimensions of food security. *FAO Rome*, 2013.

[5] W. S. Gaud. First Green Revolution: accomplishments and apprehensions. *Administrator of the Agency for International Development, Department of State, Washington, DC, USA*, 1968.

[6] Department of Economic and Social Affairs of the United Nations Secretariat. World Population Prospects The 2012 Revision. Technical report, 2013.

[7] Food and Agriculture Organization of the United Nations. FAOSTAT. *http://faostat.fao.org*, 2014.

[8] N. Alexandratos, J. Bruinsma. World agriculture towards 2030/2050: the 2012 revision. *FAO Rome*, 2012.

[9] C. Rosenzweig, J. Elliott, D. Deryng, A. C. Ruane, C. Müller, *et al.*. Assessing agricultural risks of climate change in the 21st century in a global gridded crop model intercomparison. *Proc. Natl. Acad. Sci. U.S.A.*, 111(9):3268–3273, 2014.

[10] N. Villoria, J. Elliot, H. Choi, L. Zhao. The AgMIP tool: a GEOSHARE tool for aggregating outputs from the AgMIPs global gridded crop model intercomparison project. *https://geoshareproject.org/tools/cropdatatool*, 2014.

[11] P. Pardey, J. James, J. Alston, S. Wood, B. Koo, *et al.*. Science, technology and skills. A background paper for the 2008 World Development Report of the World Bank. *The International Bank for Reconstruction and Development, The World Bank*, 2007.

[12] The World Bank. World development report 2008: agriculture for development. *The International Bank for Reconstruction and Development, The World Bank*, 2008.

[13] The World Bank. World development report 2010: Development and climate change. *The International Bank for Reconstruction and Development, The World Bank*, 2010.

[14] D. Tilman, C. Balzer, J. Hill, B. L. Befort. Global food demand and the sustainable intensification of agriculture. *Proc. Natl. Acad. Sci. U.S.A.*, 108(50):20260–20264, 2011.

[15] H. C. J. Godfray, J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, *et al.*. Food security: the challenge of feeding 9 billion people. *Science*, 327(5967):812–818, 2010.

[16] J. A. Foley, R. Defries, G. P. Asner, C. Barford, G. Bonan, *et al.*. Global consequences of land use. *Science*, 309(5734):570–574, 2005.

[17] P. L. Pingali. Green revolution: impacts, limits, and the path ahead. *Proc. Natl. Acad. Sci. U.S.A.*, 109(31):12302–12308, 2012.

[18] The Royal Society. Reaping the benefits: Science and the sustainable intensification of global agriculture. *The Royal Society*, October, 2009.

[19] J. A. Foley, N. Ramankutty, K. A. Brauman, E. S. Cassidy, J. S. Gerber, *et al.*. Solutions for a cultivated planet. *Nature*, 478(7369):337–342, 2011.

[20] D. Edwards, J. Batley. Plant genome sequencing: applications for crop improvement. *Plant Biotechnol. J.*, 8(1):2–9, 2010.

[21] M. Tester, P. Langridge. Breeding technologies to increase crop production in a changing world. *Science*, 327(5967):818–822, 2010.

[22] M. W. Bevan, C. Uauy. Genomics reveals new landscapes for crop improvement. *Genome Biol.*, 14:206, 2013.

[23] S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, *et al.*. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627–631, 2010.

[24] J. Cao, K. Schneeberger, S. Ossowski, T. Günther, S. Bender, *et al.*. Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.*, 43(10):956–963, 2011.

[25] K. Yan, P. Liu, C.-A. Wu, G.-D. Yang, R. Xu, *et al.*. Stress-induced alternative splicing provides a mechanism for the regulation of microRNA processing in Arabidopsis thaliana. *Mol. Cell.*, 48(4):521–531, 2012.

[26] S. Rasmussen, P. Barah, M. C. Suarez-Rodriguez, S. Bressendorff, P. Friis, *et al.*. Transcriptome responses to combinations of stresses in Arabidopsis thaliana. *Plant Physiol.*, 161(4):1783–1794, 2013.

[27] J. R. Harlan, D. Zohary. Distribution of wild wheats and barley. *Science*, 153(3740):1074–1079, 1966.

[28] M. Heun, R. Schäfer-Pregl, D. Klawan, R. Castagna, A. Monica, *et al.*. Site of einkorn wheat domestication identified by DNA fingerprinting. *Science*, 278(5341):1312–1314, 1997.

[29] L. Watson, M. J. Dallwitz. The grass genera of the world: descriptions, illustrations, identification, and information retrieval; including synonyms, morphology, anatomy, physiology, phytochemistry, cytology, classification, pathogens, world and local distribution, and references. *http://delta-intkey.com*, release 02, 1992.

[30] J. G. West, C. L. McIntyre, R. Appels. Evolution and systematic relationships in the Triticeae (Poaceae). *Plant Syst. Evol.*, 160(1-2):1–28, 1988.

[31] M. E. Barkworth, R. Von Bothmer. Scientific names of the Triticeae. In G. J. Muehlbaurer, C. Feuillet, editors, *Genetics and genomics of the Triticeae*, pages 3–30. 2009.

[32] B. C. Curtis, S. Rajaram, H. Gómez Macpherson. Bread wheat: improvement and production. *FAO Rome*, 2002.

[33] G. Charmet, C. Ravel, F. Balfourier. Phylogenetic analysis in the Festuca-Lolium complex using molecular markers and ITS rDNA. *Theor. Appl. Genet.*, 94(8):1038–1046, 1997.

[34] B. S. Gaut. Evolutionary dynamics of grass genomes. *New Phytol.*, 154(1):15–28, 2002.

[35] S. Bolot, M. Abrouk, U. Masood-Quraishi, N. Stein, J. Messing, *et al.*. The 'inner circle' of the cereal genomes. *Curr. Opin. Plant Biol.*, 12(2):119–125, 2009.

[36] M. Pfeifer, M. Martis, T. Asp, K. F. X. Mayer, T. Lübberstedt, *et al.*. The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics. *Plant Physiol.*, 161(2):571–582, 2013.

[37] P. F. Stevens. Angiosperm phylogeny website. *http://www.mobot.org/MOBOT/research/APweb*, version 12, 2001.

[38] E. A. Kellogg. The grasses: a case study in macroevolution. *Annu. Rev. Ecol. Syst.*, 31(1):217–238, 2000.

[39] M. D. Bennett, I. J. Leitch. Plant DNA C-values database. *http://data.kew.org/cvalues*, release 6., 2012.

[40] A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, *et al.*. The Sorghum bicolor genome and the diversification of grasses. *Nature*, 457(7229):551–556, 2009.

[41] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, *et al.*. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326(5956):1112–1115, 2009.

[42] The International Brachypodium Initiative. Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, 463(7282):763–768, 2010.

[43] J. Jia, S. Zhao, X. Kong, Y. Li, G. Zhao, *et al.*. Aegilops tauschii draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, 496(7443):91–95, 2013.

[44] H.-Q. Ling, S. Zhao, D. Liu, J. Wang, H. Sun, *et al.*. Draft genome of the wheat A-genome progenitor Triticum urartu. *Nature*, 496(7443):87–90, 2013.

[45] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*, 436(7052):793–800, 2005.

[46] B. McClintock. The significance of responses of the genome to challenge. *Science*, 226(4676):792–801, 1984.

[47] K. Song, P. Lu, K. Tang, T. C. Osborn. Rapid genome change in synthetic polyploids of Brassica and its implications for polyploid evolution. *Proc. Natl. Acad. Sci. U.S.A.*, 92(17):7719–7723, 1995.

[48] H. Ozkan, A. A. Levy, M. Feldman. Allopolyploidy-induced rapid genome evolution in the wheat (Aegilops-Triticum) group. *Plant Cell*, 13(8):1735–1747, 2001.

[49] T. Eilam, Y. Anikster, E. Millet, J. Manisterski, M. Feldman. Nuclear DNA amount and genome downsizing in natural and synthetic allopolyploids of the genera Aegilops and Triticum. *Genome*, 51(8):616–627, 2008.

[50] J.-H. Mun, S.-J. Kwon, T.-J. Yang, Y.-J. Seol, M. Jin, *et al.*. Genome-wide comparative analysis of the Brassica rapa gene space reveals genome shrinkage and differential loss of duplicated genes after whole genome triplication. *Genome Biol.*, 10(10):R111, 2009.

[51] R. B. Flavell, M. D. Bennett, J. B. Smith, D. B. Smith. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochem. Genet.*, 12(4):257–269, 1974.

[52] A. P. Tikhonov, P. J. SanMiguel, Y. Nakajima, N. M. Gorenstein, J. L. Bennetzen, *et al.*. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc. Natl. Acad. Sci. U.S.A.*, 96(13):7409–7414, 1999.

[53] N. V. Fedoroff. Transposable elements, epigenetics , and genome evolution. *Science*, 338(6108):758–767, 2012.

[54] E. A. Kellogg. Update on Evolution Evolutionary History of the Grasses 1. *Plant Physiol.*, 125(3):1198–1205, 2001.

[55] G. Moore, K. Devos, Z. Wang, M. Gale. Grasses, line up and form a circle. *Curr. Biol.*, 5(7):737–739, 1995.

[56] M. D. Gale, K. M. Devos. Comparative genetics in the grasses. *Proc. Natl. Acad. Sci. U.S.A.*, 95(5):1971–1974, 1998.

[57] K. M. Devos. Updating the 'crop circle'. *Curr. Opin. Plant Biol.*, 8(2):155–162, 2005.

[58] J. Salse, M. Abrouk, S. Bolot, N. Guilhot, E. Courcelle, *et al.*. Reconstruction of monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl. Acad. Sci. U.S.A.*, 106(35):14908–14913, 2009.

[59] J. L. Bennetzen, M. Freeling. Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genet.*, 9(8):259–261, 1993.

[60] N. Poursarebani, R. Ariyadasa, R. Zhou, D. Schulte, B. Steuernagel, *et al.*. Conserved synteny-based anchoring of the barley genome physical map. *Funct. Integr. Genomics*, 13(3):339–350, 2013.

[61] K. F. X. Mayer, S. Taudien, M. Martis, H. Simková, P. Suchánková, *et al.*. Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.*, 151(2):496–505, 2009.

[62] K. F. X. Mayer, M. Martis, P. E. Hedley, H. Simková, H. Liu, *et al.*. Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*, 23(4):1249–1263, 2011.

[63] P. Hernandez, M. Martis, G. Dorado, M. Pfeifer, S. Gálvez, *et al.*. Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.*, 69(3):377–386, 2012.

[64] P. R. Shewry. Wheat. *J. Exp. Bot.*, 60(6):1537–1553, 2009.

[65] Y. Matsuoka. Evolution of polyploid triticum wheats under cultivation: the role of domestication, natural hybridization and allopolyploid speciation in their diversification. *Plant Cell Physiol.*, 52(5):750–764, 2011.

[66] M. Nesbitt, D. Samuel. From staple crop to extinction? The archaeology and history of the hulled wheats. *Proc. 1th Int. Workshop Hulled Wheats*, 1995.

[67] G. Petersen, O. Seberg, M. Yde, K. Berthelsen. 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.*, 39(1):70–82, 2006.

[68] V. J. Nalam, M. I. Vales, C. J. W. Watson, S. F. Kianian, O. Riera-Lizarazu. Map-based analysis of genes affecting the brittle rachis character in tetraploid wheat (Triticum turgidum L.). *Theor. Appl. Genet.*, 112(2):373–381, 2006.

[69] T. Marcussen, S. R. Sandve, L. Heier, M. Spannagl, M. Pfeifer, *et al.*. Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, 345(6194):1250092, 2014.

[70] The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat genome. *Science*, 345(6194):1251788, 2014.

[71] J. S. Escobar, C. Scornavacca, A. Cenci, C. Guilhaumon, S. Santoni, *et al.*. Multigenic phylogeny and analysis of tree incongruences in Triticeae (Poaceae). *BMC Evol. Biol.*, 11:181, 2011.

[72] M. M. Martis, R. Zhou, G. Haseneyer, T. Schmutzer, J. Vrána, *et al.*. Reticulate evolution of the rye genome. *Plant Cell*, 25(10):3685–3698, 2013.

[73] P. Civáň, Z. Ivaničová, T. A. Brown. Reticulated origin of domesticated emmer wheat supports a dynamic model for the emergence of agriculture in the fertile crescent. *PloS One*, 8(11):e81955, 2013.

[74] J. Ramsey, D. W. Schemske. Pathways, mechanisms, and rates of polyploid formation in flowering plants. *Annu. Rev. Ecol. Syst.*, 29(1):467–501, 1998.

[75] Y. Van de Peer, S. Maere, A. Meyer. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.*, 10(10):725–732, 2009.

[76] J. A. Fawcett, S. Maere, Y. Van de Peer. Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event. *Proc. Natl. Acad. Sci. U.S.A.*, 106(14):5737–5742, 2009.

[77] H. Tang, X. Wang, J. E. Bowers, R. Ming, M. Alam, *et al.*. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res.*, 18(12):1944–1954, 2008.

[78] O. Jaillon, J.-M. Aury, F. Brunet, J.-L. Petit, N. Stange-Thomann, *et al.*. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. *Nature*, 431(7011):946–957, 2004.

[79] S. Kuraku, A. Meyer, S. Kuratani. Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after? *Mol. Biol. Evol.*, 26(1):47–59, 2009.

[80] D. R. Scannell, G. Butler, K. H. Wolfe. Yeast genome evolution — the origin of the species. *Yeast*, 24(11):929–942, 2007.

[81] H. Kihara, T. Ono. Chromosomenzahlen und systematische Gruppierung der Rumex-Arten. *Zeitschrift für Zellforschung und Mikroskopische Anatomie*, 4(3):475–481, 1926.

[82] L. Comai. The advantages and disadvantages of being polyploid. *Nat. Rev. Genet.*, 6(11):836–846, 2005.

[83] Z. J. Chen. Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.*, 58:377–406, 2007.

[84] J. Masterson. Stomatal size in fossile plants: evidence for polyploid in majority of Angiosperms. *Science*, 264(5157):421–424, 1994.

[85] J. E. Bowers, B. A. Chapman, J. Rong. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422(6930):433–438, 2003.

[86] B. S. Gaut. Patterns of chromosomal duplication in maize and their implications for comparative maps of the grasses. *Genome Res.*, (11):55–66, 2001.

[87] J. Yu, J. Wang, W. Lin, S. Li, H. Li, *et al.*. The genomes of Oryza sativa: a history of duplications. *PLoS Biol.*, 3(2):e38, 2005.

[88] A. H. Paterson, J. F. Wendel, H. Gundlach, H. Guo, J. Jenkins, *et al.*. Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*, 492(7429):423–427, 2012.

[89] M. Molnár-Láng, G. Linc, A. Logojan, J. Sutka. Production and meiotic pairing behaviour of new hybrids of winter wheat (Triticum aestivum) x winter barley (Hordeum vulgare). *Genome*, 43(6):1045–1054, 2000.

[90] M. Feldman, A. A. Levy. Genome evolution due to allopolyploidization in wheat. *Genetics*, 192(3):763–774, 2012.

[91] N. Al-Kaff, E. Knight, I. Bertin, T. Foote, N. Hart, *et al.*. Detailed dissection of the chromosomal region containing the Ph1 locus in wheat Triticum aestivum: with deletion mutants and expression profiling. *Ann. Bot.*, 101(6):863–872, 2008.

[92] F. K. Yousafzai, N. Al-Kaff, G. Moore. Structural and functional relationship between the Ph1 locus protein 5B2 in wheat and CDK2 in mammals. *Funct. Integr. Genomics*, 10(2):157–266, 2010.

[93] F. Salamini, H. Ozkan, A. Brandolini, R. Schäfer-Pregl, W. Martin. Genetics and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.*, 3(6):429–441, 2002.

[94] J. Dubcovsky, J. Dvorak. Genome plasticity a key factor in the success of polyploid wheat under domestication. *Science*, 316(5833):1862–1866, 2007.

[95] J. E. Melaragno, B. Mehrotra, A. W. Coleman. Relationship between endopolyploidy and cell size in epidermal tissue of Arabidopsis. *Plant Cell*, 5(11):1661–1668, 1993.

[96] T. Galitski. Ploidy regulation of gene expression. *Science*, 285(5425):251–254, 1999.

[97] R. L. Weiss, J. R. Kukora, J. Adams. The relationship between enzyme activity, cell geometry, and fitness in Saccharomyces cervisiae. *Proc. Natl. Acad. Sci. U.S.A.*, 72(3):794–798, 1975.

[98] U. C. Lavania, S. Srivastava, S. Lavania, S. Basu, N. K. Misra, *et al.*. Autopolyploidy differentially influences body size in plants, but facilitates enhanced accumulation of secondary metabolites, causing increased cytosine methylation. *Plant J.*, 71(4):539–549, 2012.

[99] S. P. Otto, J. Whitton. Polyploid incidence and evolution. *Annu. Rev. Genet.*, 34:401–437, 2000.

[100] S. P. Otto. The evolutionary consequences of polyploidy. *Cell*, 131(3):452–462, 2007.

[101] J. Wang, L. Tian, H.-S. Lee, N. E. Wei, H. Jiang, *et al.*. Genomewide nonadditive gene regulation in Arabidopsis allotetraploids. *Genetics*, 172(1):507–517, 2006.

[102] J. Wang, L. Tian, H.-S. Lee, Z. J. Chen. Nonadditive regulation of FRI and FLC loci mediates flowering-time variation in Arabidopsis allopolyploids. *Genetics*, 173(2):965–974, 2006.

[103] A. Tayalé, C. Parisod. Natural pathways to polyploidy in plants and consequences for genome reorganization. *Cytogenet. Genome Res.*, 140(2-4):79–96, 2013.

[104] V. W. Mayer, A. Aguilera. High levels of chromosome instability in polyploids of Saccharomyces cerevisiae. *Mutat. Res.*, 231(2):177–186, 1990.

[105] S. D. Ferris, G. S. Whitt. Evolution of the differential regulation of duplicate genes after polyploidization. *J. Mol. Evol.*, 12(4):267–317, 1979.

[106] H. S. Lee, Z. J. Chen. Protein-coding genes are epigenetically regulated in Arabidopsis polyploids. *Proc. Natl. Acad. Sci. U.S.A.*, 98(12):6753–6758, 2001.

[107] M. J. Hegarty, G. L. Barker, I. D. Wilson, R. J. Abbott, K. J. Edwards, *et al.*. Transcriptome shock after interspecific hybridization in senecio is ameliorated by genome duplication. *Curr. Biol.*, 16(16):1652–1659, 2006.

[108] Z. Hu, Z. Han, N. Song, L. Chai, Y. Yao, *et al.*. Epigenetic modification contributes to the expression divergence of three TaEXPA1 homoeologs in hexaploid wheat (Triticum aestivum). *New Phytol.*, 197(4):1344–1352, 2013.

[109] G. Hu, J. S. Hawkins, C. E. Grover, J. F. Wendel. The history and disposition of transposable elements in polyploid Gossypium. *Genome*, 53(8):599–607, 2010.

[110] B. C. Thomas, B. Pedersen, M. Freeling. Following tetraploidy in an Arabidopsis ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res.*, 16(7):934–946, 2006.

[111] J. C. Schnable, N. M. Springer, M. Freeling. Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl. Acad. Sci. U.S.A.*, 108(10):4069–4074, 2011.

[112] J. F. Wendel. Genome evolution in polyploids. *Plant Mol. Biol.*, 42(1):225–249, 2000.

[113] K. L. Adams, R. Cronn, R. Percifield, J. F. Wendel. Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proc. Natl. Acad. Sci. U.S.A.*, 100(8):4649–4654, 2003.

[114] L. Flagel, J. Udall, D. Nettleton, J. Wendel. Duplicate gene expression in allopolyploid Gossypium reveals two temporally distinct phases of expression evolution. *BMC Biol.*, 6:16, 2008.

[115] L. E. Flagel, L. Chen, B. Chaudhary, J. F. Wendel. Coordinated and fine-scale control of homoeologous gene expression in allotetraploid cotton. *J. Hered.*, 100(4):487–490, 2009.

[116] R. A. Rapp, J. A. Udall, J. F. Wendel. Genomic expression dominance in allopolyploids. *BMC Biol.*, 7:18, 2009.

[117] L. Comai, A. P. Tyagi, K. Winter, R. Holmes-Davis, S. H. Reynolds, *et al.*. Phenotypic instability and rapid gene silencing in newly formed arabidopsis allotetraploids. *Plant Cell*, 12(9):1551–1568, 2000.

[118] Z. Ni, E. Kim, M. Ha, E. Lackey, J. Liu, *et al.*. Altered circadian rhythms regulate growth vigor in hybrids and allopolyploids. *Nature*, 457(7227):327–331, 2009.

[119] R. J. A. Buggs, S. Chamala, W. Wu, L. Gao, G. D. May, *et al.*. Characterization of duplicate gene evolution in the recent natural allopolyploid Tragopogon miscellus by next-generation sequencing and Sequenom iPLEX MassARRAY genotyping. *Mol. Ecol.*, 19(Suppl. 1):132–146, 2010.

[120] H. Shaked, K. Kashkush, H. Ozkan, M. Feldman, A. A. Levy. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell*, 13(8):1749–1759, 2001.

[121] K. Kashkush, M. Feldman, A. A. Levy. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics*, 160(4):1651–1659, 2002.

[122] P. He, B. R. Friebe, B. S. Gill, J.-M. Zhou. Allopolyploidy alters gene expression in the highly stable hexaploid wheat. *Plant Mol. Biol.*, 52(2):401–414, 2003.

[123] K. Mochida, Y. Yamazaki, Y. Ogihara. Discrimination of homoeologous gene expression in hexaploid wheat by SNP analysis of contigs grouped from a large number of expressed sequence tags. *Mol. Genet. Genomics*, 270(5):371–377, 2003.

[124] A. Bottley, G. M. Xia, R. M. D. Koebner. Homoeologous gene silencing in hexaploid wheat. *Plant J.*, 47(6):897–906, 2006.

[125] M. Pumphrey, J. Bai, D. Laudencia-Chingcuanco, O. Anderson, B. S. Gill. Nonadditive expression of homoeologous genes is established upon polyploidization in hexaploid wheat. *Genetics*, 181(3):1147–1157, 2009.

[126] K. L. Adams, R. Percifield, J. F. Wendel. Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. *Genetics*, 168(4):2217–2226, 2004.

[127] J. Wang, L. Tian, A. Madlung, H.-S. Lee, M. Chen, *et al.*. Stochastic and epigenetic changes of gene expression in Arabidopsis polyploids. *Genetics*, 167(4):1961–1973, 2004.

[128] A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, *et al.*. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, 1999.

[129] M. Lynch, A. Force. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154(1):459–473, 2000.

[130] S. Rastogi, D. A. Liberles. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evol. Biol.*, 5:28, 2005.

[131] J. Zhang. Evolution by gene duplication: an update. *Trends Ecol. Evol.*, 18(6):292–298, 2003.

[132] B. Liu, J. F. Wendel. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol. Phylogenet. Evol.*, 29(3):365–379, 2003.

[133] M.-J. Yoo, E. Szadkowski, J. F. Wendel. Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, 110(2):171–180, 2013.

[134] F. Cheng, J. Wu, L. Fang, S. Sun, B. Liu, *et al.*. Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa. *PloS One*, 7(5):e36442, 2012.

[135] M. Feldman, A. A. Levy, T. Fahima, A. Korol. Genomic asymmetry in allopolyploid plants: wheat as a model. *J. Exp. Bot.*, 63(2):695–709, 2012.

[136] L. R. Joppa, R. G. Cantrell. Chromosomal location of genes for grain protein content of wild tetraploid wheat. *Crop Sci.*, 30:1059–1064, 1990.

[137] The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, 408(6814):796–815, 2000.

[138] F. Sanger, A. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448, 1975.

[139] F. Sanger, S. Nicklen, A. R. Coulson. DNA sequencing with chain-terminating. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, 1977.

[140] L. M. Smith, J. Z. Sanders, R. J. Kaiser, P. Hughes, C. Dodd, *et al.*. Fluorescence detection in automated DNA sequence analysis. *Nature*, 321(6071):674–679, 1986.

[141] M. L. Metzker. Emerging technologies in DNA sequencing. *Genome Res.*, 15(12):1767–1776, 2005.

[142] M. L. Metzker. Sequencing technologies - the next generation. *Nat. Rev. Genet.*, 11(1):31–46, 2010.

[143] A. H. Paterson, M. Freeling, T. Sasaki. Grains of knowledge: genomics of model cereals. *Genome Res.*, 15(12):1643–1650, 2005.

[144] S. Ossowski, K. Schneeberger, R. M. Clark, C. Lanz, N. Warthmann, *et al.*. Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome Res.*, 18(12):2024–2033, 2008.

[145] A. M. Maxam, W. Gilbert. A new method for sequencing DNA. *Proc. Natl. Acad. Sci. U.S.A.*, 24(2):560–564, 1977.

[146] H. Swerdlow, R. Gesteland. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.*, 18(6):1415–1419, 1990.

[147] M. Adams, J. Kelley, J. Gocayne, M. Dubnick, M. Polymeropoulos, *et al.*. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, 1991.

[148] L. M. Smith, S. Fung, M. W. Hunkapiller, T. J. Hunkapiller, L. E. Hood. The synthesis of oligonucleotides containing an aliphatic aio group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res.*, 13(7):2399–2412, 1985.

[149] C. A. Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Res.*, 35(18):6227–6237, 2007.

[150] The National Center for Biotechnology Information. Growth of GenBank and WGS. *http://www.ncbi.nlm.nih.gov/genbank/statistics*, 05/30/2014, 2014.

[151] J. Shendure, H. Ji. Next-generation DNA sequencing. *Nat. Biotechnol.*, 26(10):1135–1145, 2008.

[152] E. R. Mardis. A decade's perspective on DNA sequencing technology. *Nature*, 470(7333):198–203, 2011.

[153] M. Morey, A. Fernández-Marmiesse, D. Castiñeiras, J. M. Fraga, M. L. Couce, *et al.*. A glimpse into past, present, and future DNA sequencing. *Mol. Genet. Metab.*, 110(1-2):3–24, 2013.

[154] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, *et al.*. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, 2005.

[155] Illumina Inc. Illumina - Technology - Next-generation sequencing - Sequencing by Synthesis Technology. *http://www.illumina.com/technology/next-generation-sequencing/sequencing-technology.ilmn*, 07/20/2014, 2014.

[156] Illumina Inc. Illumina sequencing systems overview. *http://www.illumina.com/systems/sequencing.ilmn*, 05/31/2014, 2014.

[157] Roche Diagnostics Corporation. Roche 454 sequencing: GS FLX+ system features. *http://454.com/products/gs-flx-system/index.asp*, 31/05/2014, 2014.

[158] D. J. Munroe, T. J. R. Harris. Third-generation sequencing fireworks at Marco Island. *Nat. Biotechnol.*, 28(5):426–428, 2010.

[159] J. Clarke, H.-c. Wu, L. Jayasinghe, A. Patel, S. Reid, *et al.*. Continuous base identification for single-molecule nanopore DNA sequencing. *Nat. Nanotechnol.*, 4(4):265–270, 2009.

[160] I. Braslavsky, B. Hebert, E. Kartalov, S. R. Quake. Sequence information can be obtained from single DNA molecules. *Proc. Natl. Acad. Sci. U.S.A.*, 100(7):3960–3964, 2003.

[161] P. M. Lundquist, C. F. Zhong, P. Zhao, A. B. Tomaney, P. S. Peluso, *et al.*. Parallel confocal detection of single molecules in real time. *Opt. Lett.*, 33(9):1026–1028, 2008.

[162] J. Doležel. Mapping single chromosomes of polyploid wheat using NanoChannel arrays. *Plant and Animal Genome Conference (PAG)*, 01/14/2014, 2014.

[163] M. Baker. De novo genome assembly: what every biologist should know. *Nat. Methods*, 9(4):333–337, 2012.

[164] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. M. Jones. ABySS: a parallel assembler for short read sequence data. *Genome Res.*, 16(6):1117–1123, 2009.

[165] R. Li, Y. Li, K. Kristiansen, J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24(5):713–714, 2008.

[166] Z. Li, Y. Chen, D. Mu, J. Yuan, Y. Shi, *et al.*. Comparison of the two major classes of assembly algorithms: overlap-layout-consensus and de-bruijn-graph. *Brief Funct. Genomics.*, 11(1):25–37, 2012.

[167] J. R. Miller, S. Koren, G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–327, 2010.

[168] D. Earl, K. Bradnam, J. St John, A. Darling, D. Lin, *et al.*. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Res.*, 21(12):2224–2241, 2011.

[169] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, *et al.*. GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, 22(3):557–567, 2012.

[170] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, *et al.*. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*, 2(1):10, 2013.

[171] T. J. Treangen, S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.*, 13(1):36–46, 2012.

[172] M. G. Claros, R. Bautista, D. Guerrero-Fernández, H. Benzerki, P. Seoane, *et al.*. Why assembling plant genome sequences is so challenging. *Biology*, 1(2):439–459, 2012.

[173] C. Trapnell, S. L. Salzberg. How to map billions of short reads onto genomes. *Nat. Biotechnol.*, 27(5):455–457, 2009.

[174] B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, 2009.

[175] R. Li, C. Yu, Y. Li, T.-W. Lam, S.-M. Yiu, *et al.*. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 2009.

[176] H. Li, J. Ruan, R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18(11):1851–1858, 2008.

[177] H. Li, R. Durbin. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

[178] M. Burrows, D. Wheeler. A block sorting lossless data compression algorithm. *Digital Equipment Corporation*, Technical, 1994.

[179] M. Ruffalo, T. LaFramboise, M. Koyutürk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796, 2011.

[180] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, 5(7):621–628, 2008.

[181] C. Trapnell, B. A. Williams, G. Pertea, A. Mortazavi, G. Kwan, *et al.*. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, 28(5):511–515, 2010.

[182] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, *et al.*. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nat. Biotechnol.*, 31(1):46–53, 2013.

[183] E. T. Wang, R. Sandberg, S. Luo, I. Khrebtukova, L. Zhang, *et al.*. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[184] Y. Marquez, J. W. S. Brown, C. Simpson, A. Barta, M. Kalyna. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis. *Genome Res.*, 22(6):1184–1195, 2012.

[185] J. A. Martin, Z. Wang. Next-generation transcriptome assembly. *Nat. Rev. Genet.*, 12(10):671–682, 2011.

[186] C. Trapnell, L. Pachter, S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[187] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, B. J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.*, 40(12):1413–1415, 2008.

[188] R. P. Dilworth. A decomposition theorem for partially ordered sets. *Ann. Math.*, 51(1):161–166, 1950.

[189] G. A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, *et al.*. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, 313(5793):1596–1604, 2006.

[190] O. Jaillon, J.-M. Aury, B. Noel, A. Policriti, C. Clepet, *et al.*. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, 449(7161):463–467, 2007.

[191] The International Barley Genome Sequencing Consortium. A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491(7426):711–717, 2012.

[192] The Potato Genome Sequencing Consortium. Genome sequence and analysis of the tuber crop potato. *Nature*, 475(7355):189–195, 2011.

[193] The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485(7400):635–641, 2012.

[194] T. P. Michael, S. Jackson. The First 50 Plant Genomes. *Plant Genome*, 6(2), 2013.

[195] Z. Peng, Y. Lu, L. Li, Q. Zhao, Q. Feng, *et al.*. The draft genome of the fast-growing non-timber forest species moso bamboo (Phyllostachys heterocycla). *Nat. Genet.*, 45(4):456–461, 2013.

[196] T. Slotte, K. M. Hazzouri, J. A. Ågren, D. Koenig, F. Maumus, *et al.*. The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.*, 45(7):831–835, 2013.

[197] J. Doležel, J. Greilhuber, S. Lucrettiii, A. Meister, M. A. Lysakt, *et al.*. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann. Bot.*, 82(Suppl. A):17–26, 1998.

[198] B. S. Gill, R. Appels, A.-M. Botha-Oberholster, C. R. Buell, J. L. Bennetzen, *et al.*. A workshop report on wheat genome sequencing: International Genome Research on Wheat Consortium. *Genetics*, 168(2):1087–1096, 2004.

[199] T. Wicker, K. F. X. Mayer, H. Gundlach, M. Martis, B. Steuernagel, *et al.*. Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell*, 23(5):1706–1718, 2011.

[200] T. E. Coram, M. L. Settles, M. Wang, X. Chen. Surveying expression level polymorphism and single-feature polymorphism in near-isogenic wheat lines differing for the Yr5 stripe rust resistance locus. *Theor. Appl. Genet.*, 117(3):401–411, 2008.

[201] A. N. Bernardo, P. J. Bradbury, H. Ma, S. Hu, R. L. Bowden, *et al.*. Discovery and mapping of single feature polymorphisms in wheat using Affymetrix arrays. *BMC Genomics*, 10:251, 2009.

[202] H. Chelaifa, V. Chagué, S. Chalabi, I. Mestiri, D. Arnaud, *et al.*. Prevalence of gene expression additivity in genetically stable wheat allohexaploids. *New Phytol.*, 197(3):730–736, 2013.

[203] S. Close, S. Wanamaker, T. Close. The HarvEST database. *http://harvest.ucr.edu*, 2010.

[204] K. Mochida, T. Yoshida, T. Sakurai, Y. Ogihara, K. Shinozaki. TriFLDB: a database of clustered full-length coding sequences from Triticeae with applications to comparative grass genomics. *Plant Physiol.*, 150(3):1135–1146, 2009.

[205] J. Doležel, S. Lucretti, I. Schubert. Plant chromosome analysis and sorting by flow cytometry. *Curr. Protoc. Cytometry*, 13(3):275–309, 1994.

[206] J. Dolezel, M. Kubaláková, E. Paux, J. Bartos, C. Feuillet. Chromosome-based genomics in the cereals. *Chromosome Res.*, 15(1):51–66, 2007.

[207] J. Doležel, J. Vrána, J. Safář, J. Bartoš, M. Kubaláková, *et al.*. Chromosomes in the flow to simplify genome analysis. *Funct. Integr. Genomics*, 12(3):397–416, 2012.

[208] J. Vrana, H. Simkova, M. Kubalakova, J. Cihalikova, J. Dolezel. Flow cytometric chromosome sorting in plants: the next generation. *Methods*, 57(3):331–337, 2012.

[209] J. Safár, J. Bartos, J. Janda, A. Bellec, M. Kubaláková, *et al.*. Dissecting large and complex genomes: flow sorting and BAC cloning of individual chromosomes from bread wheat. *Plant J.*, 39(6):960–968, 2004.

[210] J. Safár, H. Simková, M. Kubaláková, J. Cíhalíková, P. Suchánková, *et al.*. Development of chromosome-specific BAC resources for genomics of bread wheat. *Cytogenet. Genome Res.*, 129(1-3):211–223, 2010.

[211] J. Breen, T. Wicker, M. Shatalina, Z. Frenkel, I. Bertin, *et al.*. A physical map of the short arm of wheat chromosome 1A. *PloS One*, 8(11):e80272, 2013.

[212] C. Silvar, D. Perovic, T. Nussbaumer, M. Spannagl, B. Usadel, *et al.*. Towards positional isolation of three quantitative trait loci conferring resistance to powdery mildew in two Spanish barley landraces. *PloS One*, 8(6):e67336, 2013.

[213] E. Sears, L. M. S. Sears. The telocentric chromosomes of common wheat. *Proc. 5th Int. Wheat Genet. Symp.*, pages 389–407, 1978.

[214] E. Paux, P. Sourdille, J. Salse, C. Saintenac, F. Choulet, *et al.*. A physical map of the 1-gigabase bread wheat chromosome 3B. *Science*, 322(5898):101–104, 2008.

[215] M.-C. Luo, Y. Q. Gu, F. M. You, K. R. Deal, Y. Ma, *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.*, 110(19):7940–7945, 2013.

[216] O.-A. Olsen. Endosperm development: cellularization and cell fate specification. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, 52:233–267, 2001.

[217] O.-A. Olsen. Nuclear endosperm development in cereals and Arabidopsis thaliana. *Plant Cell*, 16(Suppl. 2004):S214–227, 2004.

[218] S. Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic Acids Res.*, 9(13):3015–3027, 1981.

[219] T. Wicker, S. Taudien, A. Houben, B. Keller, A. Graner, *et al.*. A whole-genome snapshot of 454 sequences exposes the composition of the barley genome and provides evidence for parallel evolution of genome size in wheat and barley. *Plant J.*, 59(5):712–272, 2009.

[220] L. A. Meyers, D. A. Levin. On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206, 2006.

[221] T. Matsumoto, T. Tanaka, H. Sakai, N. Amano, H. Kanamori, *et al.*. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.*, 156(1):20–28, 2011.

[222] X. J. Min, G. Butler, R. Storms, A. Tsang. OrfPredictor: predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res.*, 33(Web Server issue):W677–680, 2005.

[223] L. Li, C. J. Stoeckert, D. S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13(9):2178–2189, 2003.

[224] N. Jiang, Z. Bao, X. Zhang, S. R. Eddy, S. R. Wessler. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, 431(7008):569–573, 2004.

[225] T. Wicker, J. P. Buchmann, B. Keller. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.*, 20(9):1229–1237, 2010.

[226] S. Kurz. The Vmatch large scale sequence analysis software. *http://www.vmatch.de*, 2011.

[227] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman. Basic Local Alignment Search Tool. *J. Mol. Biol.*, 215(3):403–410, 1990.

[228] W. Wang, H. Zheng, C. Fan, J. Li, J. Shi, *et al.*. High Rate of Chimeric Gene Origination by Retroposition in Plant Genomes. *Plant Cell*, 18(8):1791–1802, 2006.

[229] C. Fan, Y. Zhang, Y. Yu, S. Rounsley, M. Long, *et al.*. The subtelomere of Oryza sativa chromosome 3 short arm as a hot bed of new gene origination in rice. *Mol. Plant*, 1(5):839–850, 2008.

[230] Z. Zhu, Y. Zhang, M. Long. Extensive structural renovation of retrogenes in the evolution of the Populus genome. *Plant Physiol.*, 151(4):1943–1951, 2009.

[231] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *J. Am. Statist. Assoc.*, 74(368):829–836, 1979.

[232] W. S. Cleveland. LOWESS: a program for smoothing scatterplots by robust locally weighted regression. *Am. Stat.*, 35:54, 1981.

[233] S. Balzer, K. Malde, A. Lanzén, A. Sharma, I. Jonassen. Characteristics of 454 pyrosequencing data-enabling realistic simulation with flowsim. *Bioinformatics*, 26(ECCB 2010):i420–i425, 2010.

[234] W. Li, A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, 2006.

[235] D. C. Richter, F. Ott, A. F. Auch, R. Schmid, D. H. Huson. MetaSim: a sequencing simulator for genomics and metagenomics. *PloS One*, 3(10):e3373, 2008.

[236] T. Tanaka, B. A. Antonio, S. Kikuchi, T. Matsumoto, Y. Nagamura, *et al.*. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.*, 36(Database issue):D1028–1033, 2008.

[237] J. Rong, F. A. Feltus, L. Liu, L. Lin, A. H. Paterson. Gene copy number evolution during tetraploid cotton radiation. *Heredity*, 105(5):463–472, 2010.

[238] L. L. Qi, B. Echalier, S. Chao, G. R. Lazo, G. E. Butler, *et al.*. A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics*, 168(2):701–712, 2004.

[239] M. Feldman, B. Liu, G. Segal, S. Abbo, A. A. Levy, *et al.*. Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics*, 147(3):1381–1387, 1997.

[240] A. N. Massa, H. Wanjugi, K. R. Deal, K. O'Brien, F. M. You, *et al.*. Gene space dynamics during the evolution of Aegilops tauschii, Brachypodium distachyon, Oryza sativa, and Sorghum bicolor genomes. *Mol. Biol. Evol.*, 28(9):2537–2547, 2011.

[241] F. Choulet, T. Wicker, C. Rustenholz, E. Paux, J. Salse, *et al.*. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, 22(6):1686–1701, 2010.

[242] A. Haudry, A. Cenci, C. Ravel, T. Bataillon, D. Brunel, *et al.*. Grinding up wheat: a massive loss of nucleotide diversity since domestication. *Mol. Biol. Evol.*, 24(7):1506–1517, 2007.

[243] J. Dvorak, Z.-L. Yang, F. M. You, M.-C. Luo. Deletion polymorphism in wheat chromosome regions with contrasting recombination rates. *Genetics*, 168(3):1665–1675, 2004.

[244] J. O. Korbel, P. M. Kim, X. Chen, A. E. Urban, M. Snyder, *et al.*. The current excitement about copy-number variation: how it relates to gene duplication and protein families. *Curr. Opin. Struct. Biol.*, 18(3):366–374, 2009.

[245] A. H. Paterson, B. A. Chapman, J. C. Kissinger, J. E. Bowers, F. A. Feltus, *et al.*. Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon. *Trends Genet.*, 22(11):597–602, 2006.

[246] R. De Smet, K. L. Adams, K. Vandepoele, M. C. E. Van Montagu, S. Maere, *et al.*. Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants. *Proc. Natl. Acad. Sci. U.S.A.*, 110(8):2898–2903, 2013.

[247] M. Nei, I. B. Rogozin, H. Piontkivska. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc. Natl. Acad. Sci. U.S.A.*, 97(20):10866–1071, 2000.

[248] M. A. Nowak, M. C. Boerlijst, J. Cooke, J. M. Smith. Evolution of genetic redundancy. *Nature*, 388(6638):167–171, 1997.

[249] J. Zhang, Y.-P. Zhang, H. F. Rosenberg. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. *Nat. Genet.*, 30(4):411–415, 2002.

[250] T. Rattei, P. Tischler, S. Götz, M.-A. Jehl, J. Hoser, *et al.*. SIMAP - a comprehensive database of pre-calculated protein sequence similarities, domains, annotations and clusters. *Nucleic Acids Res.*, 38(Database issue):D223–226, 2010.

[251] G. P. Bolwell, K. Bozak, A. Zimmerlint. Plant Cytochrome P450. *Phytochmistry*, 37(6):1491–1506, 1994.

[252] P. Pedas, C. A. Hebbern, J. K. Schjoerring, P. E. Holm, S. Husted. Differential capacity for high-affinity manganese uptake contributes to differences between barley genotypes in tolerance to low manganese availability 1. *Plant Physiol.*, 139(11):1411–1420, 2005.

[253] X. Dai, K. Mashiguchi, Q. Chen, H. Kasahara, Y. Kamiya, *et al.*. The biochemical mechanism of auxin biosynthesis by an arabidopsis YUCCA flavin-containing monooxygenase. *J. Biol. Chem.*, 288(3):1448–1457, 2013.

[254] H. Lokstein, B. Grimm. Chlorophyll-binding proteins. *eLS. John Wiley & Sons Ltd, Chichester*, 2013.

[255] E. Pamer, P. Cresswell. Mechanisms of MHC class I-restricted antigen processing. *Annu. Rev. Immunol.*, 16:323–358, 1998.

[256] M. P. Giovanini, K. D. Saltzmann, D. P. Puthoff, M. Gonzalo, H. W. Ohm, *et al.*. A novel wheat gene encoding a putative chitin-binding lectin is associated with resistance against Hessian fly. *Mol. Plant Pathol.*, 8(1):69–82, 2007.

[257] Y. Shavrukov, P. Langridge, M. Tester. Salinity tolerance and sodium exclusion in genus Triticum. *Breeding Sci.*, 59(5):671–678, 2009.

[258] S. Wang, L. Yin, H. Tanaka, K. Tanaka, H. Tsujimoto. Wheat-Aegilops chromosome addition lines showing high iron and zinc contents in grains. *Breeding Sci.*, 61(2):189–195, 2011.

[259] M. Chamaillard, S. E. Girardin, J. Viala, D. J. Philpott. Nods, Nalps and Naip: intracellular regulators of bacterial-induced inflammation. *Cell Microbiol.*, 5(9):581–592, 2003.

[260] J. G. Turner, C. Ellis, A. Devoto. The jasmonate signal pathway. *Plant Cell*, Supplement:S153–S161, 2002.

[261] S.-W. Park, E. Kaimoyo, D. Kumar, S. Mosher, D. F. Klessig. Methyl salicylate is a critical mobile signal for plant systemic acquired resistance. *Science*, 318(5847):113–116, 2007.

[262] A. C. Vlot, P.-P. Liu, R. K. Cameron, S.-W. Park, Y. Yang, *et al.*. Identification of likely orthologs of tobacco salicylic acid-binding protein 2 and their role in systemic acquired resistance in Arabidopsis thaliana. *Plant J.*, 56(3):445–456, 2008.

[263] T. Ghelis, G. Bolbach, G. Clodic, Y. Habricot, E. Miginiac, *et al.*. Protein tyrosine kinases and protein tyrosine phosphatases are involved in abscisic acid-dependent processes in Arabidopsis seeds and suspension cells. *Plant Physiol.*, 148(3):1668–1680, 2008.

[264] N. G. Cairns, M. Pasternak, A. Wachter, C. S. Cobbett, A. J. Meyer. Maturation of Arabidopsis seeds is dependent on glutathione biosynthesis within the Embryo. *Plant Physiol.*, 141(6):446–455, 2006.

[265] S. V. Goryunova, E. M. J. Salentijn, N. N. Chikida, E. Z. Kochieva, I. M. van der Meer, *et al.*. Expansion of the gamma-gliadin gene family in Aegilops and Triticum. *BMC Evol. Biol.*, 12(1):215, 2012.

[266] L. Zhang, G. Zhao, J. Jia, X. Liu, X. Kong. Molecular characterization of 60 isolated wheat MYB genes and analysis of their expression during abiotic stress. *J. Exp. Bot.*, 63(1):203–214, 2012.

[267] C. Jacq, J. R. Miller, G. G. Brownlee. A Pseudogene laevis Structure in 5S DNA of Xenopus laevis. *Cell*, 12(September):109–120, 1977.

[268] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, *et al.*. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.

[269] Y. Liu, P. M. Harrison, V. Kunin, M. Gerstein. Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes. *Genome Biol.*, 5(9):R64, 2004.

[270] D. Zheng, A. Frankish, R. Baertsch, P. Kapranov, A. Reymond, *et al.*. Pseudogenes in the ENCODE regions: consensus annotation, analysis of transcription, and evolution. *Genome Res.*, 17(6):839–851, 2007.

[271] W. H. Li, T. Gojobori, M. Nei. Pseudogenes as a paradigm of neutral evolution. *Nature*, 292(5820):237–239, 1981.

[272] E. F. Vanin. Processed pseudogenes: characteristics and evolution. *Annu. Rev. Genet.*, 19:253–272, 1985.

[273] A. J. Mighell, N. R. Smith, P. A. Robinson, A. F. Markham. Vertebrate pseudogenes. *FEBS Lett.*, 468(2-3):109–114, 2000.

[274] E. S. Balakirev, F. J. Ayala. Pseudogenes: are they "junk" or functional DNA? *Annu. Rev. Genet.*, 37:123–151, 2003.

[275] M. Morgante, S. Brunner, G. Pea, K. Fengler, A. Zuccolo, *et al.*. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.*, 37(9):997–1002, 2005.

[276] D. Torrents, M. Suyama, E. Zdobnov, P. Bork. A genome-wide survey of human pseudogenes. *Genome Res.*, 13(12):2559–2567, 2003.

[277] K. Yamada, J. Lim, J. M. Dale, H. Chen, P. Shinn, *et al.*. Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, 302(5646):842–846, 2003.

[278] L. Poliseno, L. Salmena, J. Zhang, B. Carver, W. J. Haveman, *et al.*. A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature*, 465(7301):1033–1038, 2010.

[279] J. M. Eirín-López, L. Rebordinos, A. P. Rooney, J. Rozas. The birth-and-death evolution of multigene families revisited. *Genome Dyn.*, 7:170–196, 2012.

[280] A. Bateman, E. Birney, L. Cerruti, R. Durbin, L. Etwiller, *et al.*. The Pfam protein families database. *Nucleic Acids Res.*, 30(1):276–280, 2002.

[281] Z. Yang, R. Nielsen. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.*, 17(1):32–43, 2000.

[282] Z. Yang. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, 24(8):1586–1591, 2007.

[283] M. M. Babu, L. M. Iyer, S. Balaji, L. Aravind. The natural history of the WRKY-GCM1 zinc fingers and the relationship between transcription factors and transposons. *Nucleic Acids Res.*, 34(22):6505–6520, 2006.

[284] A. F. Bent, B. N. Kunkel, D. Dahlbeck, K. L. Brown, R. Schmidt, *et al.*. RPS2 of Arabidopsis thaliana: a leucine-rich repeat class of plant disease resistance genes. *Science*, 265(5180):1856–1860, 1994.

[285] B. C. Meyers, A. Kozik, A. Griego, H. Kuang, R. W. Michelmore. Genome-Wide Analysis of NBS-LRR – Encoding Genes in Arabidopsis. *Plant Cell*, 15(4):809–834, 2003.

[286] A. Chini, J. J. Grant, M. Seki, K. Shinozaki, G. J. Loake. Drought tolerance established by enhanced expression of the CC-NBS-LRR gene, ADR1, requires salicylic acid, EDS1 and ABI1. *Plant J.*, 38(5):810–822, 2004.

[287] C. Bai, P. Sen, K. Hofmann, L. Ma, M. Goebl, *et al.*. SKP1 connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. *Cell*, 86(2):263–274, 1996.

[288] D. Skowyra, K. L. Craig, M. Tyers, S. J. Elledge, J. W. Harper. F-box proteins are receptors that recruit phosphorylated substrates to the SCF ubiquitin-ligase complex. *Cell*, 91(2):209–219, 1997.

[289] C. Zou, M. D. Lehti-Shiu, F. Thibaud-Nissen, T. Prakash, C. R. Buell, *et al.*. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.*, 151(1):3–15, 2009.

[290] M. R. Woodhouse, B. Pedersen, M. Freeling. Transposed genes in Arabidopsis are often associated with flanking repeats. *PLoS Genet.*, 6(5):e1000949, 2010.

[291] K. Hanada, C. Zou, M. D. Lehti-Shiu, K. Shinozaki, S.-H. Shiu. Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. *Plant Physiol.*, 148(2):993–1003, 2008.

[292] M. Nei, A. P. Rooney. Concerted and birth-and-death evolution of multigene families. *Annu. Rev. Genet.*, 39:121–152, 2005.

[293] Y. Q. Gu, D. Coleman-Derr, X. Kong, O. D. Anderson. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four Triticeae genomes. *Plant Physiol.*, 135(1):459–470, 2004.

[294] R. Riley, V. Chapman. Genetic control of the cytologically diploid behaviour of hexaploid wheat. *Nature*, 182(4637):713–715, 1958.

[295] R. J. A. Buggs, S. Chamala, W. Wu, J. A. Tate, P. S. Schnable, *et al.*. Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin. *Curr. Biol.*, 22(3):248–252, 2012.

[296] J. Doležel, M. Kubaláková, J. Číhalíková, P. Suchánková, H. Simková. Chromosome analysis and sorting using flow cytometry. *Methods Mol. Biol.*, 701:221–238, 2011.

[297] T. Belova, B. Zhan, J. Wright, M. Caccamo, T. Asp, *et al.*. Integration of mate pair sequences to improve shotgun assemblies of flow-sorted chromosome arms of hexaploid wheat. *BMC Genomics*, 14:222, 2013.

[298] M. Borodovsky, K. E. Ruddl, E. V. Koonin. Intrinsic and extrinsic approaches for detecting. *Nucleic Acids Res.*, 22(22):4756–4767, 1994.

[299] R. Guigó, P. Flicek, J. F. Abril, A. Reymond, J. Lagarde, *et al.*. EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.*, 7(Suppl 1):S2, 2006.

[300] S. J. Goodswen, P. J. Kennedy, J. T. Ellis. Evaluating high-throughput ab initio gene finders to discover proteins encoded in eukaryotic pathogen genomes missed by laboratory techniques. *PloS One*, 7(11):e50609, 2012.

[301] P. Agarwal, J. F. Abril, T. Wiehe, J. W. Fickett, R. Guigo. Comparative Gene Prediction in Human and Mouse. *Genome Res.*, 13(1):108–117, 2003.

[302] M. R. Brent. How does eukaryotic gene prediction work? *Nat. Biotechnol.*, 25(8):883–885, 2007.

[303] M. Yandell, D. Ence. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, 13(5):329–342, 2012.

[304] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, *et al.*. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, 29(7):644–652, 2011.

[305] G. Gremme, V. Brendel, M. E. Sparks, S. Kurtz. Engineering a software tool for gene structure prediction in higher organisms. *Inform. Software Tech.*, 47(15):965–978, 2005.

[306] A. Roberts, H. Pimentel, C. Trapnell, L. Pachter. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics*, 27(17):2325–2329, 2011.

[307] S. A. Filichkin, H. D. Priest, S. A. Givan, R. Shen, D. W. Bryant, *et al.*. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. *Genome Res.*, 20(1):45–58, 2010.

[308] A. S. N. Reddy, Y. Marquez, M. Kalyna, A. Barta. Complexity of the alternative splicing landscape in plants. *Plant Cell*, 25(10):3657–3683, 2013.

[309] T. Lu, G. Lu, D. Fan, C. Zhu, W. Li, *et al.*. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. *Genome Res.*, 20(9):1238–1249, 2010.

[310] R. Brenchley, M. Spannagl, M. Pfeifer, G. L. A. Barker, R. D'Amore, *et al.*. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491(7426):705–710, 2012.

[311] F. Choulet, A. Alberti, S. Theil, N. Glover, V. Barbe, *et al.*. Structural and functional partitioning of bread wheat chromosome 3B. *Science*, 345(6194):1249721, 2014.

[312] A. Bottley, R. M. D. Koebner. Variation for homoeologous gene silencing in hexaploid wheat. *Plant J.*, 56(2):297–302, 2008.

[313] R. L. Tatusov, E. V. Koonin, D. J. Lipman. A genomic perspective on protein families. *Science*, 278(5338):631–637, 1997.

[314] K. M. Devos, J. Dubcovsky, J. Dvořák, C. N. Chinoy, M. D. Gale. Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theor. Appl. Genet.*, 91(2):282–288, 1995.

[315] Miftahudin, K. Ross, X.-F. Ma, A. A. Mahmoud, J. Layton, *et al.*. Analysis of expressed sequence tag loci on wheat chromosome group 4. *Genetics*, 168(2):651–663, 2004.

[316] M. Morgan, S. Falcon, R. Gentleman. GSEABase: gene set enrichment data structures and methods. *http://www.bioconductor.org/packages/release/bioc/html/GSEABase.html*.

[317] A. J. Enright, S. Van Dongen, C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, 30(7):1575–1584, 2002.

[318] R. Suzuki, H. Shimodaira. pvclust: an R package for hierarchical clustering with p-value. *http://cran.r-project.org/web/packages/pvclust*, 2011.

[319] G. Ast. How did alternative splicing evolve? *Nat. Rev. Genet.*, 5(10):773–782, 2004.

[320] I. Lamberto, R. Percudani, R. Gatti, C. Folli, S. Petrucco. Conserved alternative splicing of Arabidopsis transthyretin-like determines protein localization and S-allantoin synthesis in peroxisomes. *Plant Cell*, 22(5):1564–1574, 2010.

[321] Y. Meng, C. Shao, X. Ma, H. Wang. Introns targeted by plant microRNAs: a possible novel mechanism of gene regulation. *Rice (N. Y.)*, 6(1):8, 2013.

[322] J. C. Schöning, C. Streitner, I. M. Meyer, Y. Gao, D. Staiger. Reciprocal regulation of glycine-rich RNA-binding proteins via an interlocked feedback loop coupling alternative splicing to nonsense-mediated decay in Arabidopsis. *Nucleic Acids Res.*, 36(22):6977–6987, 2008.

[323] M. Kalyna, C. G. Simpson, N. H. Syed, D. Lewandowska, Y. Marquez, *et al.*. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis. *Nucleic Acids Res.*, 40(6):2454–2469, 2012.

[324] N. Leviatan, N. Alkan, D. Leshkowitz, R. Fluhr. Genome-wide survey of cold stress regulated alternative splicing in Arabidopsis thaliana with tiling microarray. *PloS One*, 8(6):e66511, 2013.

[325] S. Foissac, M. Sammeth. ASTALAVISTA: dynamic and flexible analysis of alternative splicing events in custom gene datasets. *Nucleic Acids Res.*, 35(Web Server issue):W297–299, 2007.

[326] B.-B. Wang, V. Brendel. Genomewide comparative analysis of alternative splicing in plants. *Proc. Natl. Acad. Sci. U.S.A.*, 103(18):7175–7180, 2006.

[327] M. Ibba, D. Söll. Quality control mechanisms during translation. *Science*, 286(5446):1893–1897, 1999.

[328] L. E. Maquat, G. G. Carmichael, N. York. Quality control of mRNA function. *Cell*, 104:173–176, 2001.

[329] Y.-F. Chang, J. S. Imam, M. F. Wilkinson. The nonsense-mediated decay RNA surveillance pathway. *Annu. Rev. Biochem.*, 76:51–74, 2007.

[330] S. Rayson, L. Arciga-Reyes, L. Wootton, M. De Torres Zabala, W. Truman, *et al.*. A Role for nonsense-mediated mRNA decay in plants: pathogen responses are induced in Arabidopsis thaliana NMD mutants. *PloS One*, 7(2):e31917, 2012.

[331] N. Riehs-Kearnan, J. Gloggnitzer, B. Dekrout, C. Jonak, K. Riha. Aberrant growth and lethality of Arabidopsis deficient in nonsense-mediated RNA decay factors is caused by autoimmune-like response. *Nucleic Acids Res.*, 40(12):5615–5624, 2012.

[332] B. P. Lewis, R. E. Green, S. E. Brenner. Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U.S.A.*, 100(1):189–192, 2003.

[333] F. Lejeune, L. E. Maquat. Mechanistic links between nonsense-mediated mRNA decay and pre-mRNA splicing in mammalian cells. *Curr. Opin. Cell Biol.*, 17(3):309–315, 2005.

[334] R. E. Green, B. P. Lewis, R. T. Hillman, M. Blanchette, L. F. Lareau, *et al.*. Widespread predicted nonsense-mediated mRNA decay of alternatively-spliced transcripts of human normal and disease genes. *Bioinformatics*, 19(Suppl. 1):i118–i121, 2003.

[335] S. Kertész, Z. Kerényi, Z. Mérai, I. Bartos, T. Pálfy, *et al.*. Both introns and long 3'-UTRs operate as cis-acting elements to trigger nonsense-mediated decay in plants. *Nucleic Acids Res.*, 34(21):6147–6157, 2006.

[336] N. H. Syed, M. Kalyna, Y. Marquez, A. Barta, J. W. S. Brown. Alternative splicing in plants-coming of age. *Trends Plant Sci.*, 17(10):616–623, 2012.

[337] The FANTOM Consortium and the RIKEN Genome Exploration Research Group Phase I & II Team. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420(6915):563–573, 2002.

[338] E. Conti, E. Izaurralde. Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. *Curr. Opin. Cell Biol.*, 17(3):316–325, 2005.

[339] Z. Zhang, D. Xin, P. Wang, L. Zhou, L. Hu, *et al.*. Noisy splicing, more than expression regulation, explains why some exons are subject to nonsense-mediated mRNA decay. *BMC Biol.*, 7:23, 2009.

[340] D. A. de Lima Morais, P. M. Harrison. Large-scale evidence for conservation of NMD candidature across mammals. *PloS One*, 5(7):e11695, 2010.

[341] N. Chantret, J. Salse, F. Sabot. Molecular basis of evolutionary events that shaped the hardness locus in diploid and polyploid wheat species (Triticum and Aegilops). *Plant Cell*, 17(4):1033–1045, 2005.

[342] A. E. Blechl, O. D. Anderson. Expression of a novel high-molecular-weight glutenin subunit gene in transgenic wheat. *Nat. Biotechnol.*, 14(7):875–879, 1996.

[343] P. Tosi, C. S. Gritsch, J. He, P. R. Shewry. Distribution of gluten proteins in bread wheat (Triticum aestivum) grain. *Ann. Bot.*, 108(1):23–35, 2011.

[344] O. D. Anderson, N. Huo, Y. Q. Gu. The gene space in wheat: the complete γ-gliadin gene family from the wheat cultivar Chinese Spring. *Funct. Integr. Genomics*, 13(2):261–273, 2013.

[345] T. K. Pellny, A. Lovegrove, J. Freeman, P. Tosi, C. G. Love, *et al.*. Cell walls of developing wheat starchy endosperm: comparison of composition and RNA-Seq transcriptome. *Plant Physiol.*, 158(2):612–627, 2012.

[346] S. A. Gillies, A. Futardo, R. J. Henry. Gene expression in the developing aleurone and starchy endosperm of wheat. *Plant Biotechnol. J.*, 10(6):668–679, 2012.

[347] C. F. Jenner. *Storage of Starch*. Springer Berlin Heidelberg, 1982.

[348] P. A. Sabelli, B. A. Larkins. The development of endosperm in grasses. *Plant Physiol.*, 149(1):14–26, 2009.

[349] F. Dupont, S. Altenbach. Molecular and biochemical impacts of environmental factors on wheat grain development and protein synthesis. *J. Cereal Sci.*, 38(2):133–146, 2003.

[350] H. G. Opsahl-Ferstad, E. Le Deunff, C. Dumas, P. M. Rogowsky. ZmEsr, a novel endosperm-specific gene expressed in a restricted region around the maize embryo. *Plant J.*, 12(1):235–246, 1997.

[351] R. D. Thompson, G. Hueros, H. A. Becker, M. Maitz. Development and functions of seed transfer cells. *Plant Sci*, 160(5):775–783, 2001.

[352] L. Hannah. Starch formation in the cereal endosperm. *Plant Cell Monogr.*, 8:179–193, 2007.

[353] P. R. Shewry, N. G. Halford. Cereal seed storage proteins: structures, properties and role in grain utilization. *J. Exp. Bot.*, 53(370):947–958, 2002.

[354] M. A. Lopes, B. A. Larkins. Endosperm origin, development, and function. *Plant Cell*, 5(10):1383–1399, 1993.

[355] H. N. Nguyen, P. A. Sabelli, B. A. Larkins. Endoreduplication and programmed cell death in the cereal endosperm. *Plant Cell Monogr.*, 8:21–43, 2007.

[356] M. F. Belmonte, R. C. Kirkbride, S. L. Stone, J. M. Pelletier, A. Q. Bui, *et al.*. Comprehensive developmental profiles of gene activity in regions and subregions of the Arabidopsis seed. *Proc. Natl. Acad. Sci. U.S.A.*, 110(5):E435–444, 2013.

[357] J. W. Walley, Z. Shen, R. Sartor, K. J. Wu, J. Osborn, *et al.*. Reconstruction of protein networks from an atlas of maize seed proteotypes. *Proc. Natl. Acad. Sci. U.S.A.*, 110(49):E4808–4817, 2013.

[358] N. Sreenivasulu, B. Usadel, A. Winter, V. Radchuk, U. Scholz, *et al.*. Barley grain maturation and germination: metabolic pathway and regulatory network commonalities and differences highlighted by new MapMan/PageMan profiling tools. *Plant Physiol.*, 146(4):1738–1758, 2008.

[359] D. L. Laudencia-Chingcuanco, B. S. Stamova, G. R. Lazo, X. Cui, O. D. Anderson. Analysis of the wheat endosperm transcriptome. *J. Appl. Genet.*, 47(4):287–302, 2006.

[360] A. W. Schreiber, M. J. Hayden, K. L. Forrest, S. L. Kong, P. Langridge. Transcriptome-scale homeolog-specific transcript assemblies of bread wheat. *BMC Genomics*, 13:492, 2012.

[361] D. Flint. Synthesis of endosperm proteins in wheat seed during maturation. *Plant Physiol.*, 56(3):381–384, 1975.

[362] J. G. Schmalstig, W. D. Hitz. Transport and metabolism of a sucrose analog (1'-Fluorosucrose) into Zea mays L. endosperm without invertase hydrolysis. *Plant Physiol.*, 85:902–905, 1987.

[363] F. Supek, M. Bošnjak, N. Škunca, T. Šmuc. REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, 6(7):e21800, 2011.

[364] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, 18(9):1509–1517, 2008.

[365] T. Griebel, B. Zacher, P. Ribeca, E. Raineri, V. Lacroix, *et al.*. Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.*, 40(20):10073–10083, 2012.

[366] C. S. Brown, P. C. Goodwin, P. K. Sorger. Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl. Acad. Sci. U.S.A.*, 98(16):8944–8949, 2001.

[367] X. H. Xu, H. Chen, Y. L. Sang, F. Wang, J. P. Ma, *et al.*. Identification of genes specifically or preferentially expressed in maize silk reveals similarity and diversity in transcript abundance of different dry stigmas. *BMC Genomics*, 13(1):294, 2012.

[368] W. H. Vensel, C. K. Tanaka, N. Cai, J. H. Wong, B. B. Buchanan, *et al.*. Developmental changes in the metabolic protein profiles of wheat endosperm. *Proteomics*, 5(6):1594–1611, 2005.

[369] A. Serna, M. Maitz, T. O'Connell, G. Santandrea, K. Thevissen, *et al.*. Maize endosperm secretes a novel antifungal protein into adjacent maternal tissue. *Plant J.*, 25(6):687–698, 2001.

[370] T. E. Young, D. R. Gallie. Analysis of programmed cell death in wheat endosperm reveals differences in endosperm development between cereals. *Plant Mol. Biol.*, 39(5):915–926, 1999.

[371] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, *et al.*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, 13(11):2498–2504, 2003.

[372] F. Y. Peng, R. J. Weselake. Gene coexpression clusters and putative regulatory elements underlying seed storage reserve accumulation in Arabidopsis. *BMC Genomics*, 12(1):286, 2011.

[373] J. Hartigan. Clustering algorithms. In *John Wiley & Sons Inc.*. 1975.

[374] J. Hartigan, M. Wong. A k-means clustering algorithm. *J. R. Stat. Soc. Series C*, 28(1):100–108, 1979.

[375] M. Maechler, P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik. cluster: cluster analysis basics and extensions. *http://cran.r-project.org/web/packages/amap*, 2013.

[376] P. J. Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20:53–65, 1987.

[377] A. Lucas. amap: another multidimensional analysis package. *http://cran.r-project.org/web/packages/amap*, 2011.

[378] K.-S. Shin, N.-J. Kwon, Y. H. Kim, H.-S. Park, G.-S. Kwon, *et al.*. Differential roles of the ChiB chitinase in autolysis and cell death of Aspergillus nidulans. *Eukaryot. Cell*, 8(5):738–746, 2009.

[379] P. Langfelder, S. Horvath. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559, 2008.

[380] B. Zhang, S. Horvath. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Molec. Biol.*, 4(1):Article 17, 2005.

[381] A. M. Yip, S. Horvath. Gene network interconnectedness and the generalized topological overlap measure. *BMC Bioinformatics*, 8:22, 2007.

[382] P. Langfelder, B. Zhang, S. Horvath. Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24(5):719–720, 2008.

[383] S. Horvath, J. Dong. Geometric interpretation of gene coexpression network analysis. *PLoS Comput. Biol.*, 4(8):e1000117, 2008.

[384] G. Csardi, T. Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Sy:1695, 2006.

[385] N. Shitsukawa, C. Tahira, K.-I. Kassai, C. Hirabayashi, T. Shimizu, *et al.*. Genetic and epigenetic alteration among three homoeologous genes of a class E MADS box gene in hexaploid wheat. *Plant Cell*, 19(6):1723–1737, 2007.

[386] E. Finnegan, C. Sheldon, F. Jardinaud. A cluster of Arabidopsis genes with a coordinate response to an environmental stimulus. *Curr. Biol.*, 14:911–916, 2004.

[387] J. de Meaux, A. Pop, T. Mitchell-Olds. Cis-regulatory evolution of chalcone-synthase expression in the genus Arabidopsis. *Genetics*, 174(4):2181–2202, 2006.

[388] J. Salse, S. Bolot, M. Throude, V. Jouffe, B. Piegu, *et al.*. Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell*, 20(1):11–24, 2008.

[389] A. Lovegrove, R. Hooley. Gibberellin and abscisic acid signalling in aleurone. *Trends Plant Sci.*, 5(3):102–110, 2000.

[390] T. Thorbjørnsen, T. Asp, K. Jørgensen, T. H. Nielsen. Starch biosynthesis from triose-phosphate in transgenic potato tubers expressing plastidic fructose-1,6-bisphosphatase. *Planta*, 214(4):616–624, 2002.

[391] I. Hara-Nishimura, T. Shimada, K. Hatano, Y. Takeuchi, M. Nishimura. Transport of storage proteins to protein storage vacuoles is mediated by large precursor-accumulating vesicles. *Plant Cell*, 10(5):825–836, 1998.

[392] A. Heyl, J. Muth, G. Santandrea, T. O'Connell, A. Serna, *et al.*. A transcript encoding a nucleic acid-binding protein specifically expressed in maize seeds. *Mol. Genet. Genomics*, 266(2):180–189, 2001.

[393] The Baking Industry Research Trust. Bake info: Gluten. *http://www.bakeinfo.co.nz/Facts/Gluten*, 07/27/2014, 2014.

[394] P. I. Payne. Genetics of wheat storage proteins and the effect of allelic variation on bread-making quality. *Annu. Rev. Plant Physiol.*, 38(1):141–153, 1987.

[395] P. R. Shewry, J. A. Napier, A. S. Tatham. Seed storage proteins: structures and biosynthesis. *Plant Cell*, 7(7):945–956, 1995.

[396] J. E. Blochet, C. Chevalier, E. Forest, E. Pebay-Peyroula, M. F. Gautier, *et al.*. Complete amino acid sequence of puroindoline, a new basic and cystine-rich protein with a unique tryptophan-rich domain, isolated from wheat endosperm by Triton X-114 phase partitioning. *FEBS Lett.*, 329(3):336–340, 1993.

[397] M.-F. Gautier, M.-E. Aleman, A. Guirao, D. Marion, P. Joudrier. Triticum aestivum puroindolines, two basic cystine-rich seed proteins: cDNA sequence analysis and developmental gene expression. *Plant Mol. Biol.*, 25(1):43–57, 1994.

[398] C. Ravel, P. Martre, I. Romeuf, M. Dardevet, R. El-Malki, *et al.*. Nucleotide polymorphism in the wheat transcriptional activator Spa influences its pattern of expression and has pleiotropic effects on grain protein composition, dough viscoelasticity, and grain hardness. *Plant Physiol.*, 151(4):2133–2144, 2009.

[399] A. De Bustos, N. Jouve. Characterisation and analysis of new HMW-glutenin alleles encoded by the Glu-R1 locus of Secale cereale. *Theor. Appl. Genet.*, 107(1):74–83, 2003.

[400] X. Zhang, D. Liu, W. Jiang, X. Guo, W. Yang, *et al.*. PCR-based isolation and identification of full-length low-molecular-weight glutenin subunit genes in bread wheat (Triticum aestivum L.). *Theor. Appl. Genet.*, 123(8):1293–1305, 2011.

[401] G. Gazzelloni, L. Gazza, N. E. Pogna. Sequencing of the Pinb-2 locus in Triticum monococcum and Triticum urartu. *Cereal Res. Commun.*, 40(1):3–13, 2012.

[402] D. Albani, M. C. U. Hammond-Kosack, C. Smith, S. Conlan, V. Colot, *et al.*. The wheat transcriptional activator SPA: a seed-specific bZIP Protein that recognizes the GCN4-like motif in the bifactorial endosperm box of prolamin genes. *Plant Cell*, 9(February):171–184, 1997.

[403] W. Zhang, P. Ciclitira, J. Messing. PacBio sequencing of gene families - a case study with wheat gluten genes. *Gene*, 533(2):541–546, 2014.

[404] O. D. Anderson, Y. Q. Gu, X. Kong, G. R. Lazo, J. Wu. The wheat omega-gliadin genes: structure and EST analysis. *Funct. Integr. Genomics*, 9(3):397–410, 2009.

[405] S. Wang, X. Shen, P. Ge, J. Li, S. Subburaj, *et al.*. Molecular characterization and dynamic expression patterns of two types of γ-gliadin genes from Aegilops and Triticum species. *Theor. Appl. Genet.*, 125(7):1371–1384, 2012.

[406] T. Sugiyama, A. Rafalski, D. Peterson, D. Söll. A wheat HMW glutenin subunit gene reveals a highly repeated structure. *Nucleic Acids Res.*, 13(24):8729–8737, 1985.

[407] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, *et al.*. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23(21):2947–2948, 2007.

[408] A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton. Jalview version 2 - a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–1191, 2009.

[409] S. Anders, P. T. Pyl, W. Huber. HTSeq A Python framework to work with high-throughput sequencing data. *bioRxiv preprint*, doi:10.110, 2014.

[410] O. D. Anderson, F. C. Greene, R. E. Yip, N. G. Halfordt, P. R. Shewryl, *et al.*. Nucleotide sequences of the two high-molecular-weight glutenin genes from the D-genome of a hexaploid bread wheat, Triticum aestivum L. cv Cheyenne. *Nucleic Acids Res.*, 17(1):461–462, 1989.

[411] J. M. Field, P. R. Shewry, B. J. Miflin. Solubilisation and characterisation of wheat gluten proteins: Correlations between the amount of aggregated proteins and baking quality. *J. Sci. Food Agric.*, 34(4):370–377, 1983.

[412] P. R. Shwry, A. S. Tatham, F. Barro, P. Barcelo, P. Lazzeri. Biotechnology of breadmaking: unraveling and manipulating the multi-protein gluten complex. *Biotechnology*, 13(11):1185–1190, 1995.

[413] F. Barro, L. Rooke, F. Békés, P. Gras, A. S. Tatham, *et al.*. Transformation of wheat with high molecular weight subunit genes results in improved functional properties. *Nat. Biotechnol.*, 15(12):1295–1299, 1997.

[414] O. D. Anderson, F. C. Greene. The characterization and comparative analysis of high-molecular-weight glutenin genes from genomes A and B of a hexaploid bread wheat. *Theor. Appl. Genet.*, 77(5):689–700, 1989.

[415] F. M. Dupont, W. H. Vensel, C. K. Tanaka, W. J. Hurkman, S. B. Altenbach. Deciphering the complexities of the wheat flour proteome using quantitative two-dimensional electrophoresis, three proteases and tandem mass spectrometry. *Proteome Sci.*, 9(1):10, 2011.

[416] R. D. Thompson, D. Bartels, N. P. Harberd, R. B. Flavell. Characterization of the multigene family coding for HMW glutenin subunits in wheat using cDNA clones. *Theor. Appl. Genet.*, 67(1):87–96, 1983.

[417] F. Chen, Z. He, D. Chen, C. Zhang, Y. Zhang, *et al.*. Influence of puroindoline alleles on milling performance and qualities of chinese noodles, steamed bread and pan bread in spring wheats. *J. Cereal Sci.*, 45(1):59–66, 2007.

[418] M. J. Giroux, C. F. Morris. Wheat grain hardness results from highly conserved mutations in the friabilin components puroindoline a and b. *Proc. Natl. Acad. Sci. U.S.A.*, 95(11):6262626–6, 1998.

[419] M.-F. Gautier, P. Cosson, A. Guirao, R. Alary, P. Joudrier. Puroindoline genes are highly conserved in diploid ancestor wheats and related species but absent in tetraploid Triticum species. *Plant Sci.*, 153(1):81–91, 2000.

[420] W. Li, L. Huang, B. S. Gill. Recurrent deletions of puroindoline genes at the grain hardness locus in four independent lineages of polyploid wheat. *Plant Physiol.*, 146(1):200–212, 2008.

[421] M. Wilkinson, Y. Wan, P. Tosi, M. Leverington, J. Snape, *et al.*. Identification and genetic mapping of variant forms of puroindoline b expressed in developing wheat grain. *J. Cereal Sci.*, 48(3):722–728, 2008.

[422] F. Chen, F. Zhang, X. Cheng, C. Morris, H. Xu, *et al.*. Association of Puroindoline b-B2 variants with grain traits, yield components and flag leaf size in bread wheat (Triticum aestivum L.) varieties of the Yellow and Huai Valleys of China. *J. Cereal Sci.*, 52(2):247–253, 2010.

[423] Y. Wan, R. L. Poole, A. K. Huttly, C. Toscano-Underwood, K. Feeney, *et al.*. Transcriptome analysis of grain development in hexaploid wheat. *BMC Genomics*, 9:121, 2008.

[424] M. J. Armstrong, V. S. Hegade, G. Robins. Advances in coeliac disease. *Curr. Opin. Gastroenterol.*, 28(2):104–112, 2012.

[425] N. L. Barbosa-Morais, M. Irimia, Q. Pan, H. Y. Xiong, S. Gueroussov, *et al.*. The evolutionary landscape of alternative splicing in vertebrate species. *Science*, 338(6114):1587–1593, 2012.

[426] J.-P. Wisniewski, P. M. Rogowsky. Vacuolar H+-translocating inorganic pyrophosphatase (Vpp1) marks partial aleurone cell fate in cereal endosperm development. *Plant Mol. Biol.*, 56(3):325–337, 2004.

[427] B. Chaudhary, L. Flagel, R. M. Stupar, J. A. Udall, N. Verma, *et al.*. Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (Gossypium). *Genetics*, 182(2):503–517, 2009.

[428] K. L. Adams. Evolution of duplicate gene expression in polyploid and hybrid plants. *J. Hered.*, 98(2):136–141, 2007.

[429] K. Nakabayashi, M. Okamoto, T. Koshiba, Y. Kamiya, E. Nambara. Genome-wide profiling of stored mRNA in Arabidopsis thaliana seed germination: epigenetic and genetic regulation of transcription in seed. *Plant J.*, 41(5):697–709, 2005.

[430] C. Rustenholz, F. Choulet, C. Laugier, J. Safár, H. Simková, *et al.*. A 3,000-loci transcription map of chromosome 3B unravels the structural and functional features of gene islands in hexaploid wheat. *Plant Physiol.*, 157(4):1596–1608, 2011.

[431] M. C. Gianibelli, O. R. Larroque, F. MacRitchie, C. W. Wrigley. Biochemical, genetic and molecular characterization of wheat glutenin and its component subunits. *Theor. Appl. Genet.*, 104(2-3):497–504, 2002.

[432] J. Massman, B. Cooper, R. Horsley, S. Neate, R. Dill-Macky, *et al.*. Genome-wide association mapping of Fusarium head blight resistance in contemporary barley breeding germplasm. *Mol. Breeding*, 27(4):439–454, 2010.

[433] K. Neumann, B. Kobiljski, S. Denčić, R. K. Varshney, A. Börner. Genome-wide association mapping: a case study in bread wheat (Triticum aestivum L.). *Mol. Breeding*, 27(1):37–58, 2010.

[434] T. Komatsuda, M. Pourkheirandish, C. He, P. Azhaguvel, H. Kanamori, *et al.*. Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc. Natl. Acad. Sci. U.S.A.*, 104(4):1424–1429, 2007.

[435] K. J. Simons, J. P. Fellers, H. N. Trick, Z. Zhang, Y.-S. Tai, *et al.*. Molecular characterization of the major wheat domestication gene Q. *Genetics*, 172(1):547–555, 2006.

[436] M. Pfeifer, K. G. Kugler, S. R. Sandve, B. Zhan, H. Rudi, *et al.*. Genome interplay in the grain transcriptome of hexaploid bread wheat. *Science*, 345(6194):1250091, 2014.

Appendix

Appendix A

Abbreviations

AL	aleurone
ATP	adenosine triphosphate
BAC	bacterial artificial chromosome
bp	basepair
BEP	Bambusoideae, Ehartoideae and Pooideae
BLAST	Basic Local Alignment Search Tool
BR	biological replicate
BWT	Burrows-Wheeler transformation
CarmA	Chromosome arm Assigner
cDNA	complementary DNA
CNV	copy number variation
CI	confidence interval
contig	contiguous sequence
DPA	days post anthesis
CSS	chromosomal survey sequence
DE	differentially expressed
DNA	deoxyribonucleic acid
EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
emPCR	emulsion PCR
ENA	European Nucleotide Archive
ESR	embryo-surrounding region
EST	expressed sequence tag
E value	Expect value

FDR	false discovery rate
fl-cDNA	full-length cDNA
FPKM	fragments per kilobase exon model per million mapped reads
Gb	gigabase pair
GH	greenhouse
Gli	gliadin
Glu	glutenin
GO	gene ontology
Ha	hardness
HC	high-confidence
HMW-Glu	high molecular weight glutenin
kbp	kilobase pair
LMW-Glu	low molecular weight glutenin
LC	low-confidence
LCS	low-confidence-supported
LCG	low-copy-number genome
Mb	megabase pair
mi	minimum overlap identity
mio	million
mRNA	messenger RNA
mya	million years ago
NCL	non-coding loci
NGS	next generation sequencing
NMD	nonsense-mediated decay
OGA	orthologous group assembly
OGR	orthologous group representative
OLC	overlap-layout-consensus
PACC	Panicoideae, Arundinoideae, Chloridoideae and Centothecoideae
PCR	polymerase chain reaction
PEG	preferentially expressed gene
Pfam	protein family
Ph1	*Pairing homoeologous 1*

pin	puroindoline
PTC	premature-termination-codon
P value	Probability value
REP	repeat associated
RNA	ribonucleic Acid
RNA-seq	RNA sequencing
RPKM	reads per kilobase exon model per million mapped reads
RUST	regulated unproductive splicing and translation
SE	starchy endosperm
SNP	single nucleotide polymorphism
SPA	storage protein activators
SRA	Sequence Read Archive
Tb	terabase
TC	transfer cells
TGS	third generation sequencing
Tp	Triticeae prototype
UK	United Kingdom
USL	unsupported loci
UTR	untranslated region
WGD	whole genome duplication
WGS	whole genome shotgun

Appendix B

Additional figures

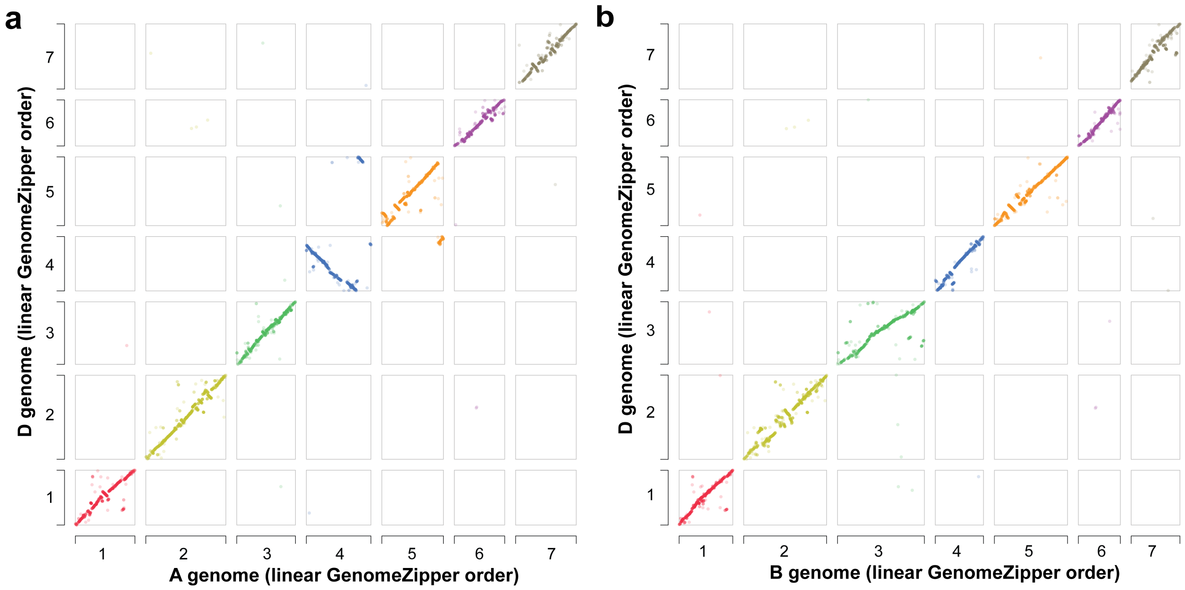


Fig. A.1. Structural analysis of homoeologous gene triplets between genomes.
To evaluate the structural representativeness of the identified homoeologous gene triplets for the entire bread wheat genome, the ordering of those genes were compared along the GenomeZippers for individual homoeologous genome (Section 4.3.3). This figures visualizes the position of genes forming single-copy homobeologous gene triplets between the **a,** A and D genomes and the **b,** B and D genomes.

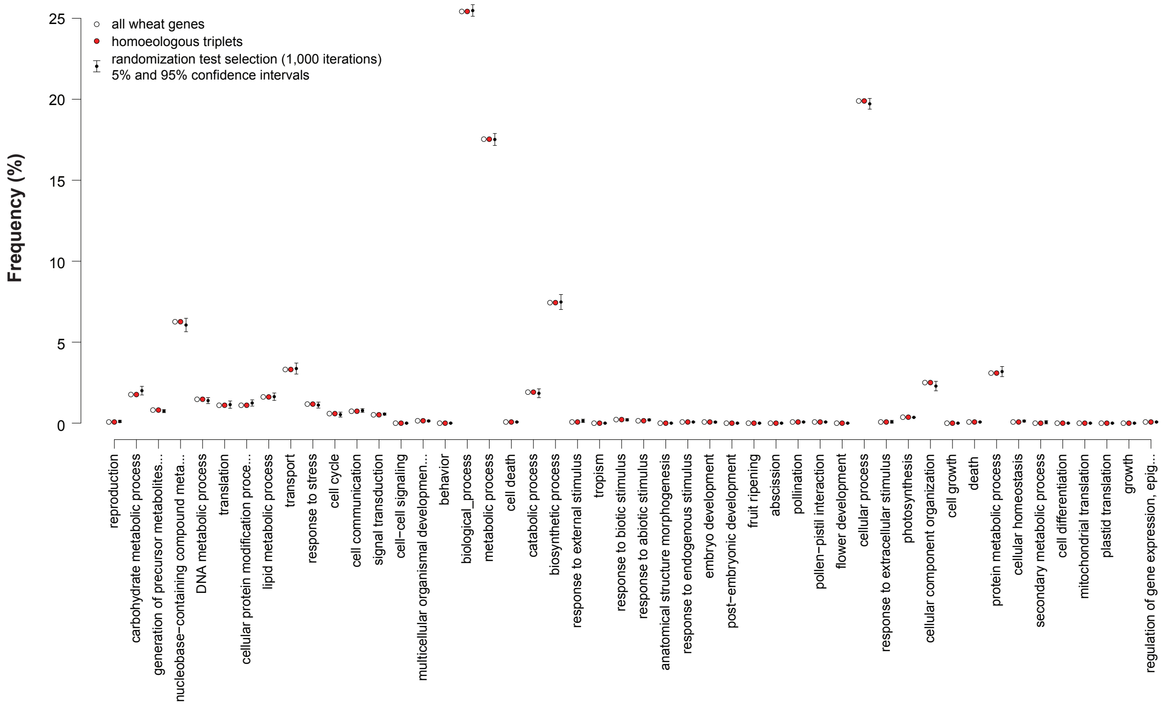


Fig. A.2. Frequency of GOSlim biological processes observed for the entire wheat genome and for homoeologous gene triplets.

To evaluate the functional representativeness of the identified homoeologous gene triplets, the GOSlim *(316)* frequency distributes for annotated GO biological process categories were compared as observed for the entire wheat gene set, for the identified homoeologous gene triplets and for sets of randomly defined gene triplets (1,000 iterations) (Section 4.3.3).

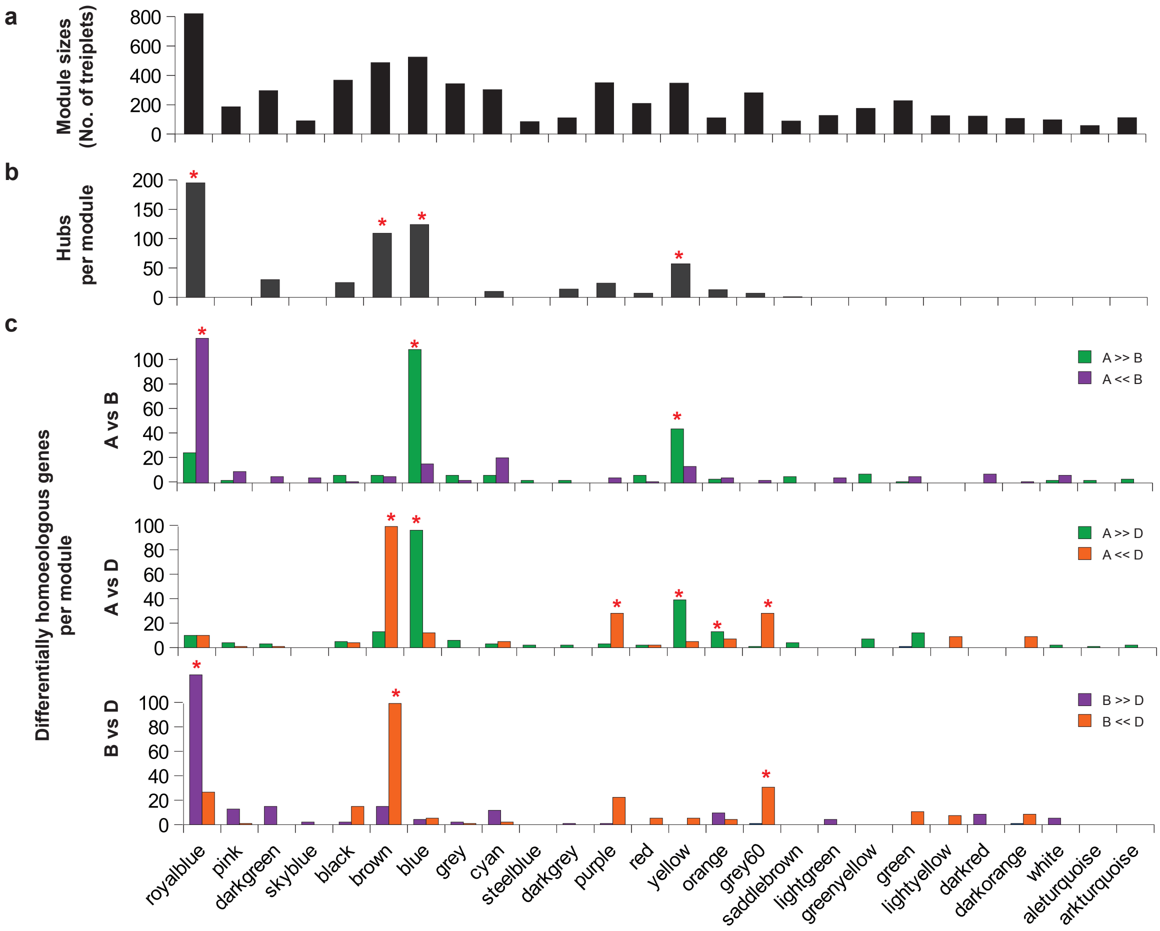
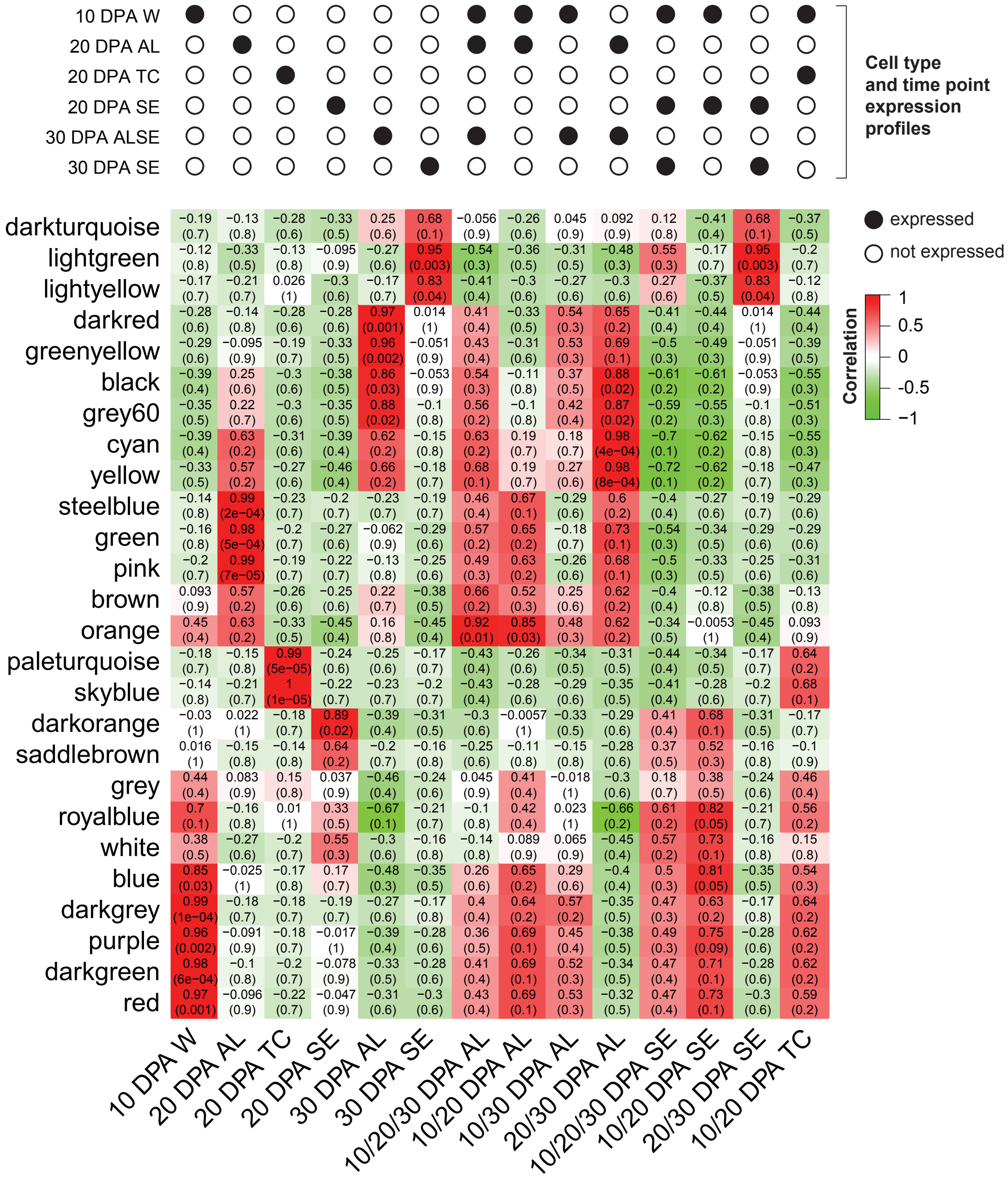


Fig. A.3. Characterization of the co-expression modules inferred for the homoeologous gene expression network.
The network-based gene expression analysis for single-copy homoeologous gene triplets revealed 25 co-expression modules, which were functionally characterized based on their expression characteristics (Section 5.5.2). **a,** Number of homoeologous gene triplets in each co-expression module. **b,** Number of identified hub genes placed into each co-expression modules. **c,** Number of differentially expressed homoeologous genes located in each co-expression module. Red stars mark co-experssion modules with a significant number of hubs (**b**) or differentially expressed homoeologous genes (**c**) (one sided Fisher's exact test with Bonferroni corrected *P* value <0.01).



10 DPA W	●	○	○	○	○	○	●	●	●	○	●	●	○	●
20 DPA AL	○	●	○	○	○	○	●	●	○	●	○	○	○	○
20 DPA TC	○	○	●	○	○	○	○	○	○	○	○	○	○	●
20 DPA SE	○	○	○	●	○	○	○	○	○	○	●	●	●	○
30 DPA ALSE	○	○	○	○	●	○	●	○	●	●	○	○	○	○
30 DPA SE	○	○	○	○	○	●	○	○	○	○	●	○	●	○

	10 DPA W	20 DPA AL	20 DPA TC	20 DPA SE	30 DPA AL	30 DPA SE	10/20/30 DPA AL	10/20 DPA AL	10/30 DPA AL	20/30 DPA AL	10/20/30 DPA SE	10/20 DPA SE	20/30 DPA SE	10/20 DPA TC
darkturquoise	−0.19 (0.7)	−0.13 (0.8)	−0.28 (0.6)	−0.33 (0.5)	0.25 (0.6)	0.68 (0.1)	−0.056 (0.9)	−0.26 (0.6)	0.045 (0.9)	0.092 (0.9)	0.12 (0.8)	−0.41 (0.4)	0.68 (0.1)	−0.37 (0.5)
lightgreen	−0.12 (0.8)	−0.33 (0.5)	−0.13 (0.8)	−0.095 (0.9)	−0.27 (0.6)	0.95 (0.003)	−0.54 (0.3)	−0.36 (0.5)	−0.31 (0.5)	−0.48 (0.3)	0.55 (0.3)	−0.17 (0.7)	0.95 (0.003)	−0.2 (0.7)
lightyellow	−0.17 (0.7)	−0.21 (0.7)	0.026 (1)	−0.3 (0.6)	−0.17 (0.7)	0.83 (0.04)	−0.41 (0.4)	−0.3 (0.6)	−0.27 (0.6)	−0.3 (0.6)	0.27 (0.6)	−0.37 (0.5)	0.83 (0.04)	−0.12 (0.8)
darkred	−0.28 (0.6)	−0.14 (0.8)	−0.28 (0.6)	−0.28 (0.6)	0.97 (0.001)	0.014 (1)	0.41 (0.4)	−0.33 (0.5)	0.54 (0.3)	0.65 (0.2)	−0.41 (0.4)	−0.44 (0.4)	0.014 (1)	−0.44 (0.4)
greenyellow	−0.29 (0.6)	−0.095 (0.9)	−0.19 (0.7)	−0.33 (0.5)	0.96 (0.002)	−0.051 (0.9)	0.43 (0.4)	−0.31 (0.6)	0.53 (0.3)	0.69 (0.1)	−0.5 (0.3)	−0.49 (0.3)	−0.051 (0.9)	−0.39 (0.5)
black	−0.39 (0.4)	0.25 (0.6)	−0.3 (0.6)	−0.38 (0.5)	0.86 (0.03)	−0.053 (0.9)	0.54 (0.3)	−0.11 (0.8)	0.37 (0.5)	0.88 (0.02)	−0.61 (0.2)	−0.61 (0.2)	−0.053 (0.9)	−0.55 (0.3)
grey60	−0.35 (0.5)	0.22 (0.7)	−0.3 (0.6)	−0.35 (0.5)	0.88 (0.02)	−0.1 (0.8)	0.56 (0.2)	−0.1 (0.8)	0.42 (0.4)	0.87 (0.02)	−0.59 (0.2)	−0.55 (0.3)	−0.1 (0.8)	−0.51 (0.3)
cyan	−0.39 (0.4)	0.63 (0.2)	−0.31 (0.6)	−0.39 (0.4)	0.62 (0.2)	−0.15 (0.8)	0.63 (0.2)	0.19 (0.7)	0.18 (0.7)	0.98 (4e−04)	−0.7 (0.1)	−0.62 (0.2)	−0.15 (0.8)	−0.55 (0.3)
yellow	−0.33 (0.5)	0.57 (0.2)	−0.27 (0.6)	−0.46 (0.4)	0.66 (0.2)	−0.18 (0.7)	0.68 (0.1)	0.19 (0.7)	0.27 (0.6)	0.98 (8e−04)	−0.72 (0.1)	−0.62 (0.2)	−0.18 (0.7)	−0.47 (0.3)
steelblue	−0.14 (0.8)	0.99 (2e−04)	−0.23 (0.7)	−0.2 (0.7)	−0.23 (0.7)	−0.19 (0.7)	0.46 (0.4)	0.67 (0.1)	−0.29 (0.6)	0.6 (0.2)	−0.4 (0.4)	−0.27 (0.6)	−0.19 (0.7)	−0.29 (0.6)
green	−0.16 (0.8)	0.98 (5e−04)	−0.2 (0.7)	−0.27 (0.6)	−0.062 (0.9)	−0.29 (0.6)	0.57 (0.2)	0.65 (0.2)	−0.18 (0.7)	0.73 (0.1)	−0.54 (0.3)	−0.34 (0.5)	−0.29 (0.6)	−0.29 (0.6)
pink	−0.2 (0.7)	0.99 (7e−05)	−0.19 (0.7)	−0.22 (0.7)	−0.13 (0.8)	−0.25 (0.6)	0.49 (0.3)	0.63 (0.2)	−0.26 (0.6)	0.68 (0.1)	−0.5 (0.3)	−0.33 (0.5)	−0.25 (0.6)	−0.31 (0.6)
brown	0.093 (0.9)	0.57 (0.2)	−0.26 (0.6)	−0.25 (0.6)	0.22 (0.7)	−0.38 (0.5)	0.66 (0.2)	0.52 (0.3)	0.25 (0.6)	0.62 (0.2)	−0.4 (0.4)	−0.12 (0.8)	−0.38 (0.5)	−0.13 (0.8)
orange	0.45 (0.4)	0.63 (0.2)	−0.33 (0.5)	−0.45 (0.4)	0.16 (0.8)	−0.45 (0.4)	0.92 (0.01)	0.85 (0.03)	0.48 (0.3)	0.62 (0.2)	−0.34 (0.5)	−0.0053 (1)	−0.45 (0.4)	0.093 (0.9)
paleturquoise	−0.18 (0.7)	−0.15 (0.8)	0.99 (5e−05)	−0.24 (0.6)	−0.25 (0.6)	−0.17 (0.7)	−0.43 (0.4)	−0.26 (0.6)	−0.34 (0.5)	−0.31 (0.5)	−0.44 (0.4)	−0.34 (0.5)	−0.17 (0.7)	0.64 (0.2)
skyblue	−0.14 (0.8)	−0.21 (0.7)	1 (1e−05)	−0.22 (0.7)	−0.23 (0.7)	−0.2 (0.7)	−0.43 (0.4)	−0.28 (0.6)	−0.29 (0.6)	−0.35 (0.5)	−0.41 (0.4)	−0.28 (0.6)	−0.2 (0.7)	0.68 (0.1)
darkorange	−0.03 (1)	0.022 (1)	−0.18 (0.7)	0.89 (0.02)	−0.39 (0.4)	−0.31 (0.5)	−0.3 (0.6)	−0.0057 (1)	−0.33 (0.5)	−0.29 (0.6)	0.41 (0.4)	0.68 (0.1)	−0.31 (0.5)	−0.17 (0.7)
saddlebrown	0.016 (1)	−0.15 (0.8)	−0.14 (0.8)	0.64 (0.2)	−0.2 (0.7)	−0.16 (0.8)	−0.25 (0.6)	−0.11 (0.8)	−0.15 (0.8)	−0.28 (0.6)	0.37 (0.5)	0.52 (0.3)	−0.16 (0.8)	−0.1 (0.9)
grey	0.44 (0.4)	0.083 (0.9)	0.15 (0.8)	0.037 (0.9)	−0.46 (0.4)	−0.24 (0.6)	0.045 (0.9)	0.41 (0.4)	−0.018 (1)	−0.3 (0.6)	0.18 (0.7)	0.38 (0.5)	−0.24 (0.6)	0.46 (0.4)
royalblue	0.7 (0.1)	−0.16 (0.8)	0.01 (1)	0.33 (0.5)	−0.67 (0.1)	−0.21 (0.7)	−0.1 (0.8)	0.42 (0.4)	0.023 (1)	−0.66 (0.2)	0.61 (0.2)	0.82 (0.05)	−0.21 (0.7)	0.56 (0.2)
white	0.38 (0.5)	−0.27 (0.6)	−0.2 (0.7)	0.55 (0.3)	−0.3 (0.6)	−0.16 (0.8)	−0.14 (0.8)	0.089 (0.9)	0.065 (0.9)	−0.45 (0.4)	0.57 (0.2)	0.73 (0.1)	−0.16 (0.8)	0.15 (0.8)
blue	0.85 (0.03)	−0.025 (1)	−0.17 (0.8)	0.17 (0.7)	−0.48 (0.3)	−0.35 (0.5)	0.26 (0.6)	0.65 (0.2)	0.29 (0.6)	−0.4 (0.4)	0.5 (0.3)	0.81 (0.05)	−0.35 (0.5)	0.54 (0.3)
darkgrey	0.99 (1e−04)	−0.18 (0.7)	−0.18 (0.7)	−0.19 (0.7)	−0.27 (0.6)	−0.17 (0.8)	0.4 (0.4)	0.64 (0.2)	0.57 (0.2)	−0.35 (0.5)	0.47 (0.3)	0.63 (0.2)	−0.17 (0.8)	0.64 (0.2)
purple	0.96 (0.002)	−0.091 (0.9)	−0.18 (0.7)	−0.017 (1)	−0.39 (0.4)	−0.28 (0.6)	0.36 (0.5)	0.69 (0.1)	0.45 (0.4)	−0.38 (0.5)	0.49 (0.3)	0.75 (0.09)	−0.28 (0.6)	0.62 (0.2)
darkgreen	0.98 (6e−04)	−0.1 (0.8)	−0.2 (0.7)	−0.078 (0.9)	−0.33 (0.5)	−0.28 (0.6)	0.41 (0.4)	0.69 (0.1)	0.52 (0.3)	−0.34 (0.5)	0.47 (0.4)	0.71 (0.1)	−0.28 (0.6)	0.62 (0.2)
red	0.97 (0.001)	−0.096 (0.9)	−0.22 (0.7)	−0.047 (0.9)	−0.31 (0.6)	−0.3 (0.6)	0.43 (0.4)	0.69 (0.1)	0.53 (0.3)	−0.32 (0.5)	0.47 (0.4)	0.73 (0.1)	−0.3 (0.6)	0.59 (0.2)

Fig. A.4. Correlation of co-expression module eigengenes with pre-defined cell type and time point expression profiles.

To investigate cell-type and time point specificity of the identified co-expression modules (Section 5.5.2), the corresponding eigengene vectors were correlated against pre-defined expression profiles. Upper values and the heat map color intensity correspond to the measured Pearson's correlation coefficients. Values in brackets denote significance for the observed correlation (*P* value).

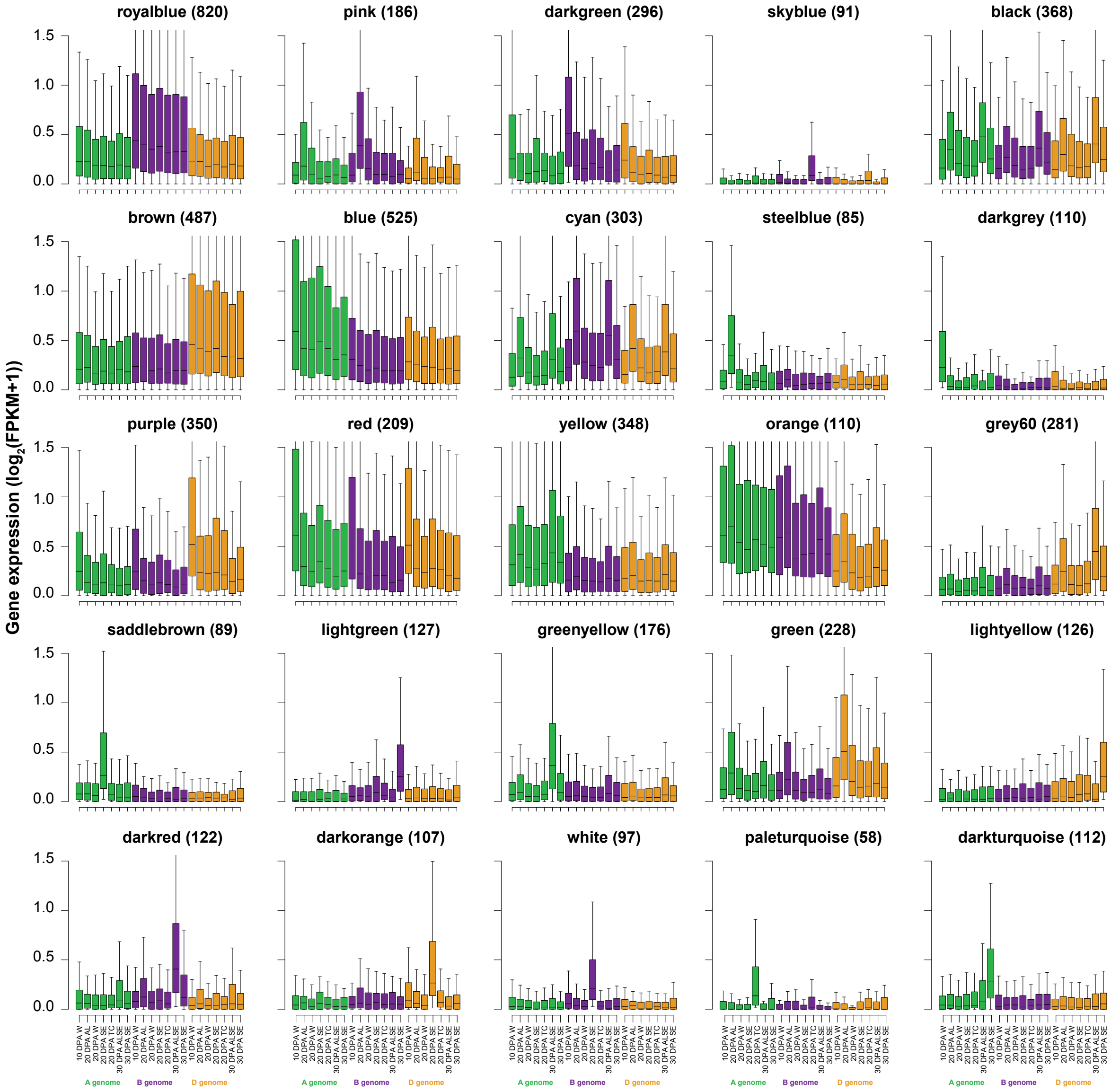


Fig. A.5. Gene expression profiles of the identified co-expression modules for the homoeologous gene expression network.
The boxplots visualize the gene expression distribution for each of the 25 identified co-expression modules identified by network-based cluster analysis of gene expression for single-copy homoeologous gene triplets (Section 5.5.2).

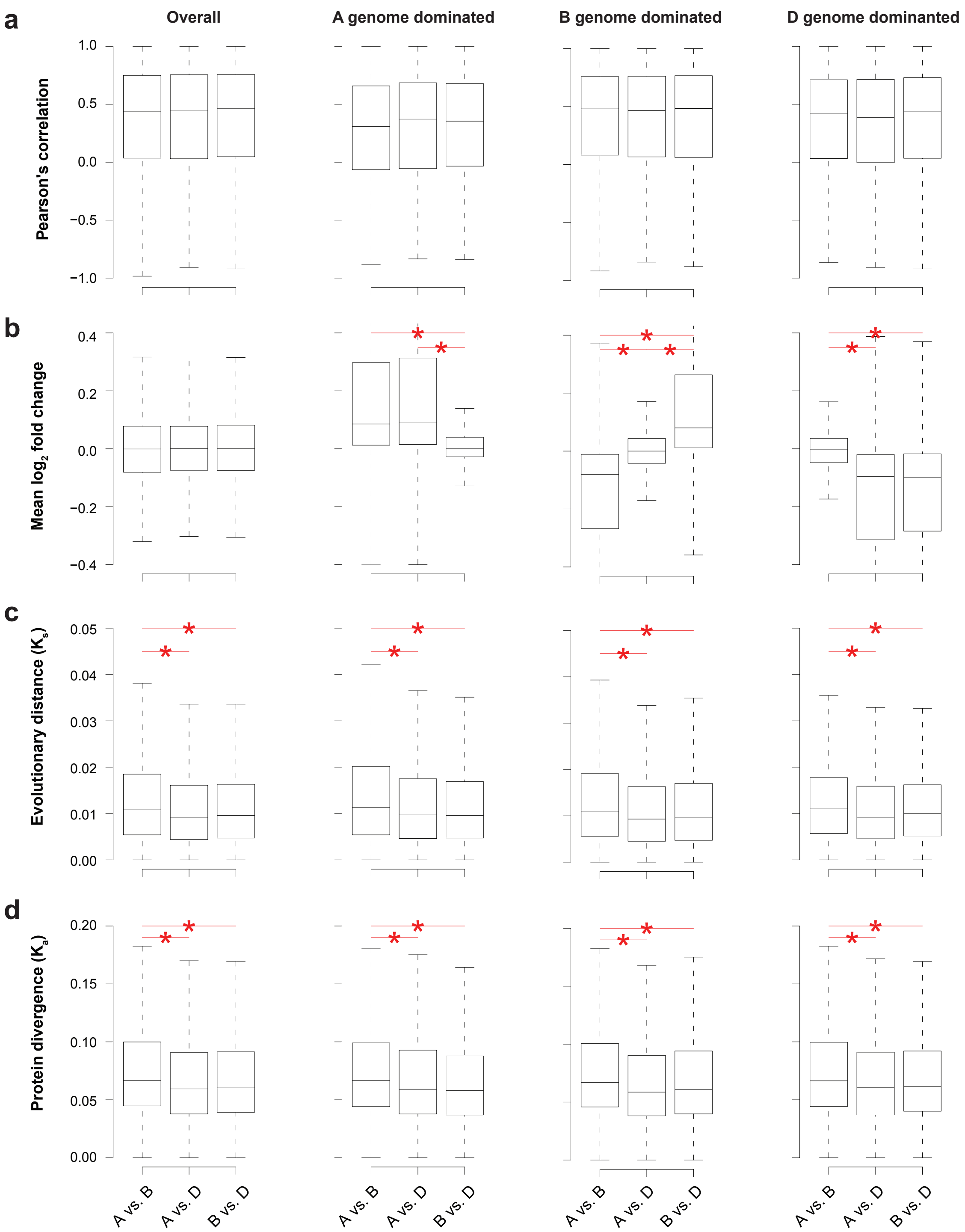


Fig. A.6. Distribution of gene expression correlation, gene expression level dominance and sequence divergence in for homoeologous triplets.

To elucidate relationship between asymmetric gene expression and sequence divergence, transcription-based features [correlation in gene expression and differences in gene expression levels (**a** and **b**)] were compared with sequence-based features [evolutionary distance and protein divergence (**c** and **d**)] for all homoeologous triplets (”overall”) as well as genome-dominated co-expression groups (Section 5.5.3). **a,** Correlation in gene expression measured by Pearson's correlation coefficient of expression values. **b,** Log$_2$ fold-changes averaged over all endosperm samples. **c,** Evolutionary distances measured by the number of synonymous substitutions per synonymous site (K_s). **d,** Protein divergence measured by the number of non-synonymous substitutions per non-synonymous site (K_a). Significant differences between distributions are marked by red stars [Wilcoxon-Mann-Whitney-Test (*P* <0.01)].

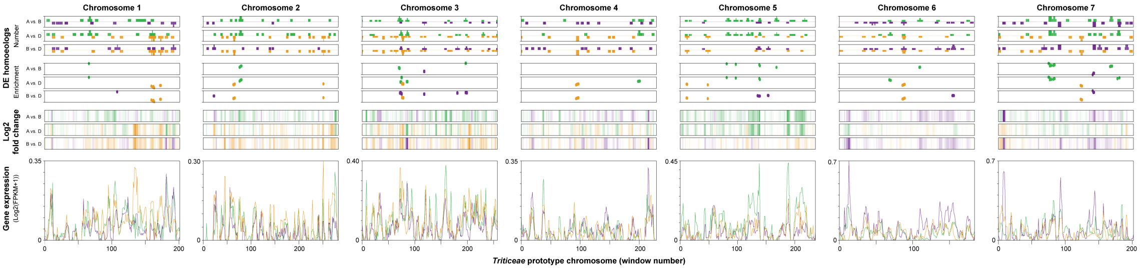


Fig. A.7. Distribution of gene expression levels for 10 DPA W along the Tp chromosomes. To investigate influence of chromosomal positioning on transcript abundances, the chromosomal distribution of gene expression was monitored along the seven Triticeae prototype (Tp) chromosomes and pairwise compared between the three homoeologous wheat genomes (A vs B, A vs D and B vs D) (Section 5.6). Therefore, a sliding window algorithm was applied calculating the median gene expression along the chromosomes (window size 50 Tp loci; window shift 10 Tp loci). For each window the top three panels count the number of significant differentially expressed (DE) homoeologous triplets between the A and B genomes, A and D genomes and B and D genomes, respectively (Section 5.5). Chromosomal segments that were significantly enriched for DE homoeologous genes were also identified and visualized by dots in the following three panels (Fisher's exact test with P value ≤ 0.05). The heat maps show the pairwise log$_2$-fold change of median gene expression between two windows, whereupon increased color intensity mark higher fold change towards one genome. The last panel show the median gene expression for each window. In all panels the A genome is colored green, the B genome purple and the D genome orange, respectively.

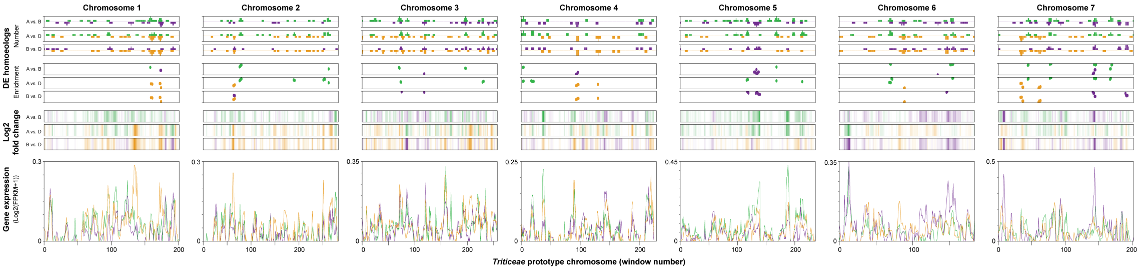


Fig. A.8. Distribution of gene expression levels for 20 DPA W along the Tp chromosomes. For description see legend of Fig. A.7.

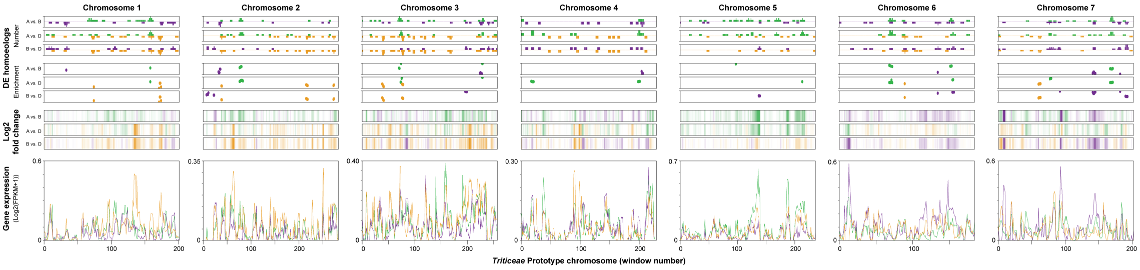


Fig. A.9. Distribution of gene expression levels for 20 DPA AL along the Tp chromosomes. For description see legend of Fig. A.7.

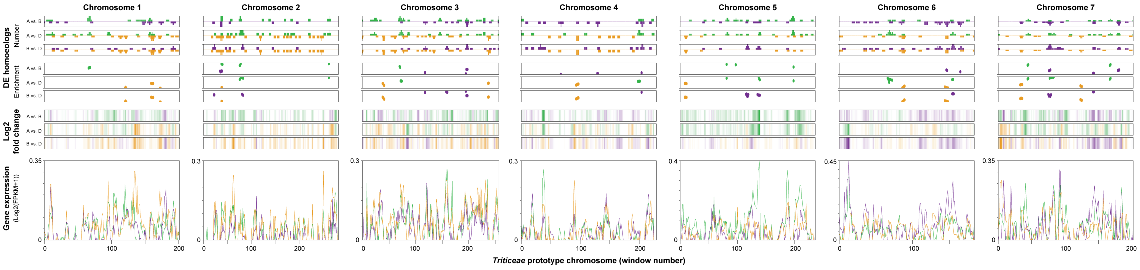


Fig. A.10. Distribution of gene expression levels for 20 DPA SE along the Tp chromosomes. For description see legend of Fig. A.7.

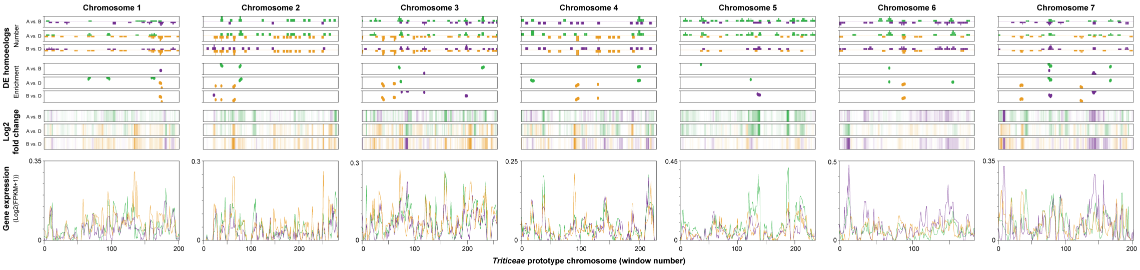


Fig. A.11. Distribution of gene expression levels for 20 DPA TC along the Tp chromosomes. For description see legend of Fig. A.7.

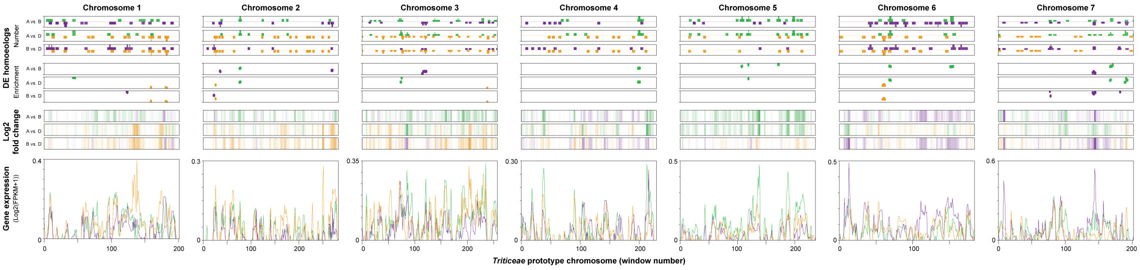


Fig. A.12. Distribution of gene expression levels for 30 DPA TC along the Tp chromosomes. For description see legend of Fig. A.7.

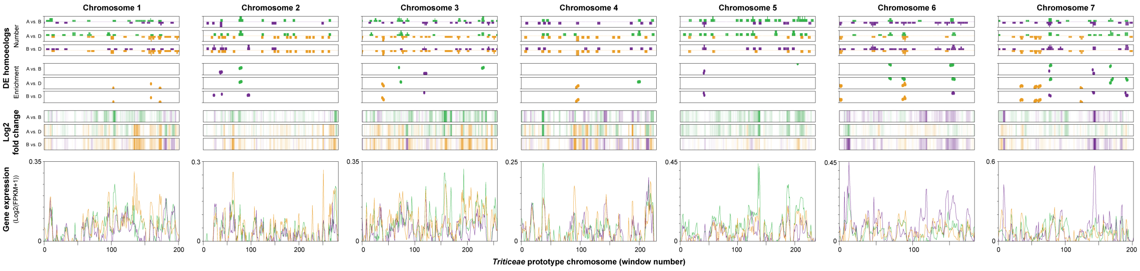


Fig. A.13. Distribution of gene expression levels for 30 DPA SE along the Tp chromosomes. For description see legend of Fig. A.7.

Appendix C

Additional tables

Table A.1. Functional enrichment analysis of preferentially expressed genes (PEGs).
To functional interpret the identified preferentially expressed genes, GO enrichment tests were performed for PEGs defined in individual endosperm cell types and developmental stages (Section 5.3.2). This table is part of Pfeifer *et al. (436)* and available as Excel file on *Science* Online (Table S1):
http://www.sciencemag.org/content/345/6194/1250091/suppl/DC1

Table A.2. Functional enrichments for individual k-means co-expression clusters.
Each identified k-means co-expression cluster was subject to GO enrichment analysis to identify over-represented functional molecular functions and biological processes related to the commonly grouped genes (Section 5.4.1). This table is part of Pfeifer *et al. (436)* and available as Excel file on *Science* Online (Table S2):
http://www.sciencemag.org/content/345/6194/1250091/suppl/DC1

Table A.3. Expression transitions between homoeologs of the A and B genomes.
For all identified homoeologous gene triplets assignment to co-expression clusters were analysed and the number of observed expression transitions, i.e. different assignment to co-expression clusters for homoeologous gene pairs were determined (Section 5.4.3). This table shows the number of A (rows) and B (columns) homoeologs that were placed into the same (diagonal) or into different co-expression clusters.

	-[(a)]	**0**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
-[(a)]	771	66	58	19	53	41	42	44	56
0	60	1840	187	251	100	85	33	133	87
I	51	150	220	58	34	21	17	22	16
II	29	199	29	77	29	42	11	21	17
III	42	106	43	32	30	18	7	12	28
IV	28	120	17	37	18	78	9	24	12
V	25	31	13	12	21	8	32	12	4
VI	58	177	20	20	16	21	7	162	22
VII	39	96	21	21	21	11	7	27	42

[(a)] No gene expression observed.

Table A.4. Expression transitions between homoeologs of the A and D genomes.
See Table A.3 for the general description. Here, the number of A (rows) and D (columns) homoeologs that were placed into the same (diagonal) or into different co-expression clusters are shown.

	-[a]	**0**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
-[a]	787	50	58	26	44	39	32	61	53
0	48	1822	169	237	98	107	28	171	96
I	61	154	202	59	41	18	21	18	15
II	29	205	35	83	34	22	11	17	18
III	34	99	37	32	44	14	17	18	23
IV	28	110	25	32	18	84	8	19	19
V	33	29	15	10	9	10	29	6	17
VI	59	168	17	18	16	24	13	152	36
VII	44	89	25	15	14	14	14	29	41

[a] No gene expression observed.

Table A.5. Expression transitions between homoeologs of the B and D genomes.
See Table A.3 for the general description. Here, the number of B (rows) and D (columns) homoeologs that were placed into the same (diagonal) or into different co-expression clusters are shown.

	-[a]	**0**	**I**	**II**	**III**	**IV**	**V**	**VI**	**VII**
-[a]	791	44	53	25	41	26	24	49	50
0	56	1804	173	237	102	105	42	177	89
I	58	166	217	51	40	20	15	19	22
II	23	252	38	97	32	40	15	13	17
III	40	100	31	29	47	17	16	18	24
IV	32	104	15	33	7	77	10	25	22
V	42	20	18	9	17	7	27	6	19
VI	42	143	14	18	16	22	15	155	32
VII	39	93	24	13	16	18	9	29	43

[a] No gene expression observed.

Table A.6. Number of aggregated transitions of homoeologous genes between the identified k-means co-expression clusters and significance tests.
The number of transitions between co-expression clusters with endosperm-specific expression profiles (cluster I to VII) were aggregated and tested for significance. Each cell counts the number of transitions observed among pairs of homoeologous genes between co-expression clusters (Tables A.3 to A.5 and Section 5.4.3). Bonferroni adjusted *P* values are given in parenthesis calculated with an one-sided Fisher's exact test. Bold numbers indicate a significant number of observed transitions (*P* <0.05). Within-cluster transitions and transitions with cluster 0 were not included in the analysis.

	I	**II**	**III**	**IV**	**V**	**VI**	**VII**
I	-	**168 (<0.001)**	**115 (<0.001)**	59 (1.00)	53 (1.00)	59 (1.00)	53 (1.00)
II	102 (0.004)	-	**95 (0.003**)	**104 (<0.001)**	37 (1.00)	51 (1.00)	52 (1.00)
III	**111 (<0.001)**	93 (1.00)	-	49 (1.00)	40 (1.00)	48 (1.00)	75 (0.275)
IV	57 (1.00)	**102 (<0.001)**	43 (1.00)	-	27 (1.00)	**68 (0.005)**	53 (1.00)
V	46 (1.00)	31 (1.00)	47 (0.173)	25 (1.00)	-	24 (1.00)	40 (1.00)
VI	51 (1.00)	56 (1.00)	48 (1.00)	**67 (0.021)**	35 (1.00)	-	**90 (<0.001)**
VII	70 (0.542)	49 (1.00)	51 (1.00)	43 (1.00)	30 (1.00)	**85 (<0.001)**	-

Table A.7. Number of significant differentially homoeologous genes grouped in co-expression modules identified by the network-based analysis of homoeologous gene expression.
This table lists the number of significant differentially expressed (DE) homoeologs ($P \leq 0.05$) identified for each co-expression module in pairwise comparisons of gene expression level between the A, B and D genomes (Section 5.5). Numbers in brackets show Bonferroni corrected P values (one-sided Fisher's exact test) and bold indicate a significant enrichment ($P \leq 0.001$).

Group	Triplets	A vs. B		A vs. D		B vs. D		Total
		A	B	A	D	B	D	
1	715	0 (1.00)	**108 (<0.001)**	7 (1.00)	16 (1.00)	**94 (<0.001)**	4 (1.00)	**124 (<0.001)**
2	649	0 (1.00)	4 (1.00)	0 (1.00)	**63 (<0.001)**	1 (1.00)	**54 (<0.001)**	70 (1.00)
3	511	**53 (<0.001)**	0 (1.00)	**46 (<0.001)**	0 (1.00)	0 (1.00)	4 (1.00)	55 (1.00)
4	505	**94 (<0.001)**	0 (1.00)	**93 (<0.001)**	0 (1.00)	8 (1.00)	3 (1.00)	**104 (<0.001)**
5	470	4 (1.00)	2 (1.00)	2 (1.00)	9 (1.00)	1 (1.00)	20 (1.00)	26 (1.00)
6	378	1 (1.00)	**38 (<0.001)**	2 (1.00)	14 (1.00)	19 (0.960)	2 (1.00)	45 (1.00)
7	350	4 (1.00)	3 (1.00)	1 (1.00) (1.00)	5 (1.00)	4 (1.00)	8 (1.00)	16 (1.00)
8	338	**30 (<0.001)**	1 (1.00)	**45 (<0.001)**	0 (1.00)	17 (1.00)	3 (1.00)	49 (1.00)
9	287	**31 (<0.001)**	0 (1.00)	6 (1.00)	7 (1.00)	0 (1.00)	26	35 (1.00)
10	271	4 (1.00)	1 (1.00)	4 (1.00)	0 (1.00)	2 (1.00)	1 (1.00)	7 (1.00)
11	249	0 (1.00)	10 (1.00)	6 (1.00)	3 (1.00)	19 (0.013)	0 (1.00)	25 (1.00)
12	231	0 (1.00)	**38 (<0.001)**	0 (1.00)	3 (1.00)	**33 (<0.001)**	0 (1.00)	42 (0.075)
13	215	2 (1.00)	2 (1.00)	1 (1.00)	**28 (<0.001)**	0 (1.00)	**24 (<0.001)**	33 (1.00)
14	184	1 (1.00)	6 (1.00)	0 (1.00)	**71 (<0.001)**	0 (1.00)	**70 (<0.001)**	**75 (<0.001)**
15	161	5 (1.00)	0 (1.00)	3 (1.00)	1 (1.00)	0 (1.00)	1 (1.00)	5 (1.00)
16	142	0 (1.00)	5 (1.00)	0 (1.00)	0 (1.00)	5 (1.00)	0 (1.00)	6 (1.00)

Table A.8. Functional enrichments for co-expression modules inferred for the homoeologous gene expression network.
Each identified co-expression module was subject to GO enrichment analysis to identify over-represented functional molecular functions and biological processes related to the commonly grouped triplets (Section 5.5.1). This table is part of Pfeifer *et al. (436)* and available as Excel file on *Science* Online (Table S3): http://www.sciencemag.org/content/345/6194/1250091/suppl/DC1