



Wissenschaftszentrum Weihenstephan für
Ernährung, Landnutzung und Umwelt
Lehrstuhl für Genomorientierte Bioinformatik

GenomeZipper - a bioinformatic approach to unravel highly complex cereal genomes

Mihaela-Maria Martis

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

Die Dissertation wurde am 23.03.2015 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 27.05.2015 angenommen.

Abstract

Cereals belong to one of the most economically important plant families (*Poaceae*). They provide half of all the calories consumed by humans, are indispensable for animal feed and the beverage industry, and have a huge potential to be used for biofuel production.

Until recently, most whole genome sequencing projects were restricted to small plant genomes. Large and complex grass genomes, including those of barley (*Hordeum vulgare*), rye (*Secale cereale*), and wheat (*Triticum aestivum*), were limited by computational requirements in sequencing, genome assembly, and identification of protein coding regions. The new next-generation sequencing technologies enabled to tackle these larger genomes. However, the assemblies are still highly fragmented and lack positional ordering.

This thesis describes a new tool termed *GenomeZipper*, which aims to identify, order and structure chromosomal survey sequences of large grass genomes that lack physical maps by exploiting the widely conserved synteny among grasses. Additionally, subsequent comparative analyses performed on the *GenomeZipper* results are described.

Applied on barley (*Hordeum vulgare*) and rye (*Secale cereale*) the *GenomeZipper* concept allowed to arrange and orient 68% and 72% of the estimated gene space, respectively. The resulting linear ordered gene maps were used to compare the conserved gene structure and organisation in these grasses. The comparison revealed several major chromosomal rearrangements in the rye genome and molecular characteristics gave evidence for introgressive hybridisation. This indicates that the rye genome evolution and speciation was influenced by reticulate evolution.

Another structural characteristic of the rye genome studied in this thesis are supernumerary B chromosomes (Bs). Bs are dispensable extra chromosomes to the normal chromosome complement (A chromosomes), which exhibit non-Mendelian inheritance. This work highlights the intraspecific origin of Bs by recombination of two A chromosomes (3R and 7R) and give insights into their evolution and composition. Large amount of organellar DNA, B specific repeats and large numbers of gene fragments appear to be characteristic for the B chromosomes of rye.

Zusammenfassung

Süßgräser (*Poaceae*) sind eine weltweit verbreitete und wirtschaftlich bedeutende Pflanzenfamilie, zu denen auch Getreide gehören. Sie liefern über die Hälfte der Kalorien für die Welternährung und sind als Futterpflanze und als Rohstoff für die Brauindustrie unverzichtbar. Dazu gewinnen sie zunehmend an Bedeutung im Bereich der Biokraftstoffproduktion.

Auf Grund von technischen und finanziellen Einschränkungen wurden zunächst nur Pflanzen mit relativ kleinen Genomen zur Sequenzierung ausgewählt. Die Sequenzierung, Assemblierung und Gendetektion in großen und komplexen Gräsergenomen, wie Gerste (*Hordeum vulgare*), Roggen (*Secale cereale*) und Weizen (*Triticum aestivum*), waren bisher durch die großen notwendigen Rechenleistungen eingeschränkt. Durch den Einsatz neuer Next-Generation-Sequencing-Technologien wurde ein Durchbruch erzielt und die Analyse solch großer Genome möglich. Dennoch sind diese Genome noch unvollständig und die Abfolge ihrer Sequenzen weitgehendst unbekannt.

Im Rahmen dieser Arbeit wird eine neue Methode vorgestellt, die als *GenomeZipper* bezeichnet wird. Das *GenomeZipper*-Konzept nutzt die konservierte Syntanie zwischen Gräsern aus, um Gensequenzen von Grassgenomen, ohne physikalische Karte, entlang von Chromosomen anzuordnen. Zusätzlich werden die Ergebnisse der vergleichenden Analysen, welche auf den vom *GenomeZipper* erstellten virtuellen Genkarten beruhen, beschrieben. Das *GenomeZipper*-Verfahren wurde auf unterschiedliche Gräser, wie Gerste und Roggen, angewandt, wobei für respektive 68% und 72% der geschätzten Gene die Reihenfolge entlang der Chromosomen bestimmt werden konnte. Der Vergleich der daraus entstandenen virtuellen Gerste- und Roggengenkarten enthüllte wesentliche chromosomale Änderungen der Anordnung im Roggen genom. Diese deuten darauf hin, dass die Roggen-Artbildung und -Evolution durch introgressive Hybridisierung und retikuläre Evolution sowie einer Reihe von Genomduplikationen beeinflusst wurde.

Ein weiterer Bestandteil dieser Arbeit ist die Analyse von B Chromosomen (Bs) in Roggen. Bs sind überzählige, entbehrliche Chromosomen, die unabhängig von den Mendelschen Gesetzen vererbt werden. Die Ergebnisse dieser Arbeit zeigen den intraspezifischen Ursprung der Roggen B Chromosomen durch Fusion zweier A Chromosomen (3R und 7R) und geben Einblicke in die Evolution sowie die Zusammensetzung der B Chromosomen. Ansammlungen von Organellen-DNA, B-spezifischen repetitiven Sequenzen und Genfragmenten scheinen charakteristisch für die B Chromosomen in Roggen zu sein.

Acknowledgements

First of all, I wish to thank my supervisors at the Institute of Bioinformatics and Systems Biology, Dr. Klaus Mayer and Prof. Dr. Hans-Werner Mewes, for giving me the opportunity to pursue a PhD. I especially appreciate the support and guidance given by Dr. Klaus Mayer concerning my work at the Plant Genome and Systems Biology group as well as the chance to travel and present my research at international conferences.

Special thanks go to Dr. Andreas Houben for introducing me to the fascinating research area on B chromosomes. Thank you for your patience, support, and comments regarding the rye B chromosomes and for the opportunity to present my results on the 3rd B-Chromosome Conference.

Thanks to everyone at PGSB for the good teamwork and constructive feedback, for lunch breaks, and for supporting my chili and tomato breeding attempt. It has been a great honor and pleasure to collaborate with you on different projects. To Heidrun for her input, feedback, and support over the last years, for Christmas baking, and for skiing and sleigh rides.

To all my friends and family, for good times, encouragements, and shoulders to cry on. To Hanna and Anna, for both exhaustive and last-minute proofreading. Without you I would have lost the day against English grammar.

My warmest thanks go to Bernhard. Thank you for supporting and encouraging me through a challenging time. I would have never ever finished my thesis without you. It is good to have you by my side.

Abbreviations

Ae	<i>Aegilops</i>
AFLP	a mplified f ragment- l ength p olymorphism
BAC	b acterial a rtificial c hromosomes
BBH	b i- d irectional B last h it
Bd	<i>Brachypodium distachyon</i>
BLAST	b asic l ocal a lignment s earch t ool
bp	b ase p airs
BP	b efore p resent
cal	c alibrated
CSS	c hromosomal s urvey s equences
DNA	d eoxyribonucleic a cid
EST	e xpressed s equence t ag
FHIT	f ragile h istidine t riad g ene
FISH	f luorescence <i>in situ</i> h ybridisation
fl-cDNA	f ull-length c omplementary D N A
Gbp	g iga b ase p airs
GSS	g enome s urvey s equences
IRAP	i nter- r etrotransposon a mplified p olymorphism
ITS	i nternal t ranscribed s pacers
ISSR	i nter- s imple s equence r epeat
kbp	k ilo b ase p airs
LTR	l ong t erminal r epeat
Mbp	m ega b ase p airs
Mya	m illion y ears a go
NGS	n ext g eneration s equencing
NOR	n uclear o rganising r egions
Os	<i>Oryza sativa</i> (rice)
PSR	p aternal- s ex- r atio
qPCR	q uantitative p olymerase c hain r eaction
rDNA	r ibosomal D N A
RFLP	r estriction f ragment l ength p olymorphism
RNA	r ibonucleic a cid
Sb	<i>Sorghum bicolor</i> (sorghum)
Sc	<i>Secale cereale</i> (rye)
sSMC	s mall s upernumerary m arker c hromosomes
TE	t ransposable e lements

Abbreviations

WCS	whole ch romosome s hotgun
WGS	whole g enome s hotgun
WGD	whole g enome d uplication
YAML	yet a nother m ulticolumn l ayout (a human-readable data serialization standard)

List of publications

The following peer-reviewed publications are included in this thesis:

1. Mayer KF, Taudien S, **Martis M**, Šimková H, Suchánková P, Gundlach H, Wicker T, Petzold A, Felder M, Steuernagel B, Scholz U, Graner A, Platzer M, Doležel J and Stein N. *Gene content and virtual gene order of barley chromosome 1H*. **Plant Physiol.** 2009 Oct, 151(2):496-505. doi:10.1104/pp.109.142612.
2. Mayer KF, **Martis M**, Hedley PE, Šimková H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, Kubaláková M, Suchánková P, Murat F, Felder M, Nussbaumer T, Graner A, Salse J, Endo T, Sakai H, Tanaka T, Itoh T, Sato K, Platzer M, Matsumoto T, Scholz U, Doležel J, Waugh R and Stein N. *Unlocking the barley genome by chromosomal and comparative genomics*. **Plant Cell.** 2011 Apr, 23(4):1249-63. doi: 10.1105/tpc.110.082537.
3. **Martis MM**, Klemme S, Banaei-Moghaddam AM, Blattner FR, Macas J, Schmutzer T, Scholz U, Gundlach H, Wicker T, Šimková H, Novák P, Neumann P, Kubaláková M, Bauer E, Haseneyer G, Fuchs J, Doležel J, Stein N, Mayer KF and Houben A. *Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences*. **PNAS.** 2012 Aug, 14;109(33):13343-6. doi:10.1073/pnas.1204237109.
4. **Martis MM**, Zhou R, Haseneyer G, Schmutzer T, Vrána J, Kubaláková M, König S, Kugler KG, Scholz U, Hackauf B, Korzun V, Schön CC, Doležel J, Bauer E, Mayer KFX and Stein N. *Reticulate evolution of the rye (*Secale cereale L.*) genome*. **Plant Cell.** 2013 Oct, 25(10):3685-98. doi:10.1105/tpc.113.114553.

Further publications:

1. Helguera M, Rivarola M, Clavijo B, **Martis MM**, Vanzetti LS, Gonzales S, Garbus I, Leroy P, Šimková H, Valárik M, Caccamo M, Doležel J, Mayer KFX, Feuillet C, Tranquilli G, Paniego N and Echenique VC. *New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing*. **Plant Science**, accepted and in press.
2. Banaei-Moghaddam AM, **Martis MM**, Macas J, Gundlach H, and Houben A. *Genes on B chromosomes: old questions revisited with new tools*. **Biochimica et Biophysica Acta.** 2015, 1849(1):64-70. doi:10.1016/j.bbagr.2014.11.007
3. The International Wheat Genome Sequencing Consortium (IWGSC). *A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome*. **Science.** 2014 Jul, Vol. 345, no. 6194. doi:10.1126/science.1251788

List of publications

4. Kopecký D, **Martis MM**, Číhalíková J, Hřibová E, Vrána J, Bartoš J, Kopecká J, Cattonaro F, Stočes Š, Novák P, Neumann P, Macas J, Šimková H, Studer B, Asp T, Baird JH, Navrátil P, Karafiátová M, Kubaláková M, Šafář J, Mayer KF, and Doležel J. *Flow sorting and sequencing meadow fescue chromosome 4F*. **Plant Physiol.** 2013 Nov, 163(3):1323-37. PMID:24096412.
5. Lüpken T, Stein N, Perovic D, Habekuß A, Serfling A, Krämer I, Hähnel U, Steuernagel B, Scholz U, Ariyadasa R, **Martis M**, Mayer KFX, Niks RE, Collins NC, Friedt W, and Ordon F. *High-resolution mapping of the barley Ryd3 locus controlling tolerance to BYDV*. **Molecular Breeding** 2013 Oct, 1-12. doi:10.1007/s11032-013-9966-1.
6. Spannagl M, **Martis MM**, Pfeifer M, Nussbaumer T, and Mayer KFX. *Analysing complex Triticeae genomes – concepts and strategies*. **Plant Methods.** 2013 Sep, 9(1):35. doi:10.1186/1746-4811-9-35.
7. Poursarebani N, Ariyadasa R, Zhou R, Schulte D, Steuernagel B, **Martis MM**, Graner A, Schweizer P, Scholz U, Mayer K, and Stein N. *Conserved synteny-based anchoring of the barley genome physical map*. **Funct Integr Genomics.** 2013 Aug;13(3):339-50. doi:10.1007/s10142-013-0327-2.
8. Philippe R, Paux E, Bertin I, Sourdille P, Choulet F, Laugier C, Šimková H, Šafář J, Bellec A, Vautrin S, Frenkel Z, Cattonaro F, Magni F, Scalabrin S, **Martis MM**, Mayer KF, Korol A, Bergès H, Doležel J, and Feuillet C. *A high density physical map of chromosome 1BL supports evolutionary studies, map-based cloning and sequencing in wheat*. **Genome Biol.** 2013 Jun, 14(6):R64. PMID:23800011
9. Luo MC, Gu YQ, You FM, Deal KR, Ma Y, Hu Y, Huo N, Wang Y, Wang J, Chen S, Jorgensen CM, Zhang Y, McGuire PE, Pasternak S, Stein JC, Ware D, Kramer M, McCombie WR, Kianian SF, **Martis MM**, Mayer KF, Sehgal SK, Li W, Gill BS, Bevan MW, Šimková H, Doležel J, Weining S, Lazo GR, Anderson OD, and Dvorak J. *A 4-gigabase physical map unlocks the structure and evolution of the complex genome of Aegilops tauschii, the wheat D-genome progenitor*. **PNAS.** 2013 May, 110(19):7940-5. doi:10.1073/pnas.1219082110.
10. Lüpken T, Stein N, Perovic D, Habekuss A, Krämer I, Hähnel U, Steuernagel B, Scholz U, Zhou R, Ariyadasa R, Taudien S, Platzer M, **Martis M**, Mayer K, Friedt W, and Ordon F. *Genomics-based high-resolution mapping of the BaMMV/BaYMV resistance gene rym11 in barley (Hordeum vulgare L.)*. **Theor Appl Genet.** 2013 May, 126(5):1201-12. doi:10.1007/s00122-013-2047-3.
11. Pfeifer M, **Martis M**, Asp T, Mayer KF, Lübberstedt T, Byrne S, Frei U, and Studer B. *The perennial ryegrass GenomeZipper: targeted use of genome resources for comparative grass genomics*. **Plant Physiol.** 2013 Feb, 161(2):571-82. doi:10.1104/pp.112.207282.
12. Nussbaumer T, **Martis MM**, Rößner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, and Spannagl M. *MIPS PlantsDB: a database framework for comparative plant genome research*. **Nucleic Acids Res.** 2013 Jan, 41(Database issue):D1144-51. doi:10.1093/nar/gks1153.

13. International Barley Genome Sequencing Consortium. *A physical, genetic and functional sequence assembly of the barley genome*. **Nature**. 2012 Nov 29, 491(7426):711-6. doi:10.1038/nature11543.
14. Hernandez P, **Martis M**, Dorado G, Pfeifer M, Gálvez S, Schaaf S, Jouve N, Šimková H, Valárik M, Doležel J, and Mayer KFX. *Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content*. **Plant J**. 2012 Feb, 69(3):377-86. doi:10.1111/j.1365-313X.2011.04808.x.
15. Wicker T, Mayer KF, Gundlach H, **Martis M**, Steuernagel B, Scholz U, Šimková H, Kubaláková M, Choulet F, Taudien S, Platzer M, Feuillet C, Fahima T, Budak H, Doležel J, Keller B, and Stein N. *Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives*. **Plant Cell**. 2011 May, 23(5):1706-18. doi:10.1105/tpc.111.086629.
16. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, **Martis M**, Narechania A, Otiillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, Freeling M, Gingle AR, Hash CT, Keller B, Klein P, Kresovich S, McCann MC, Ming R, Peterson DG, Mehboob-ur-Rahman, Ware D, Westhoff P, Mayer KFX, Messing J, and Rokhsar DS. *The Sorghum bicolor genome and the diversification of grasses*. **Nature**. 2009 Jan, 457:551-556. doi:10.1038/nature07723.

Contents

Acknowledgements	i
Abbreviations	iii
List of publications	v
1. Introduction	1
1.1. Plant genomes	1
1.1.1. Plant genome organisation	2
1.1.2. Plant genome sequencing and assembly	5
1.1.3. Taxonomy, origin, and economic importance of grasses	6
1.1.4. Comparative genomics in grasses	10
1.2. The <i>GenomeZipper</i> approach	12
1.2.1. Motivation	12
1.2.2. The concept	13
1.3. B chromosomes	14
1.3.1. Motivation	14
1.3.2. Definition of B chromosomes	15
1.3.3. Origin of B chromosomes	16
1.3.4. B chromosome effects	18
1.3.5. Molecular composition and organisation of B chromosomes	19
1.3.6. Maintenance of B chromosomes	22
1.3.7. The B chromosomes of rye	24
2. Material and methods	27
2.1. Repeat masking	28
2.2. Identify conserved linkage blocks among grasses	28
2.3. <i>GenomeZipper</i>	28
3. Publication summaries	33
3.1. <i>Gene content and virtual gene order of barley chromosome 1H</i> Mayer K.F., Taudien S., <u>Martis M.</u> , Šimková H., Suchánková P., Gundlach H., Wicker T., Petzold A., Felder M., Steuernagl B., Scholz U., Graner A., Platzer M., Doležel J. and Stein N. 2009	33

Contents

3.2. <i>Unlocking the barley genome by chromosomal and comparative genomics</i> Mayer K.F., <u>Martis M.</u> , Hedley P.E., Šimková H., Liu H., Morris J.A., Steuernagl B., Taudien S., Roessner S., Gundlach H., Kubaláková M., Suchánková P., Murat F., Felder M., Nussbaumer T., Graner A., Salse J., Endo T., Sakai H., Tanaka T., Itoh T., Sato K., Platzer M., Matsumoto T., Scholz U., Doležel J., Waugh R. and Stein N. 2011	35
3.3. <i>Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences</i> <u>Martiset al.</u> 2012	36
3.4. <i>Reticulate evolution of the rye (<i>Secale cereale L.</i>) genome</i> <u>Martiset al.</u> 2013	37
4. Discussion and conclusions	39
4.1. Cereal genomes unlocked using the <i>GenomeZipper</i> method	39
4.2. Insights into the structure, organisation, and evolution of cereal genomes	43
4.2.1. The barley and rye genomes	45
4.2.2. Deciphering the rye B chromosome	49
A. Configuration Files	53
B. Original publications	59
Bibliography	105
Curriculum Vitae	133

1. Introduction

1.1. Plant genomes

Plants are versatile and essential components of the world's ecosystem. They maintain the environmental balance by stabilising soils, producing oxygen, and assimilating CO₂. Furthermore, they sustain the existence of all living beings by providing them with habitats and resources for food, medicines, fibre, and fuel [1]. Within the plant kingdom, plants can be assigned to two categories: the seedless plants (mosses, ferns, horsetails, and algae) and the seed-bearing plants (angiosperms and gymnosperms). Also, angiosperms (flowering plants) and gymnosperms (conifers, ginkgoes, and cycads) can be separated into two classes, the monocotyledon (monocots) and the dicotyledon (eudicots). The members of the two classes differ by leaves form, number of cotyledons and petals, secondary growth, and stems organisation. Monocots do not have secondary growth and exhibit leaves with parallel veins, one cotyledon, and flower petals as a multiple of three [2]. In contrast, eudicots are characterised by two cotyledons, by branching stems and leaf veins, and have flower petals as a multiple of four or five [2]. Most members of the flowering plants (75%) [3] belong to the dicotyledons and include economically important families like the *Fabaceae* family (soybean, beans), the *Solanaceae* family (potato, tomato, tobacco), the *Rosaceae* family (raspberries, strawberries, apples, cherries, peaches), and the *Brassicaceae* family (cauliflower, mustard). The remaining flowering plants (25%) belong to the monocotyledons, which comprise among others the large families of the *Orchidaceae* and the grasses (*Poaceae*) [3, 4].

Nowadays, over 370,000 plant species are known, among them roughly 7,000 species are being exploited as resources for human consumption and therefore have a great economic importance [5]. The plant species reveal an exceptional diversity in ecology, size, shape, structure, and longevity, as well as ingenious biochemical defence and adaptation mechanisms against pathogens and environmental changes. Hence, they include annual, biennial, perennial herbs, aquatic plants, shrubs and trees. Individuals can reach assorted sizes from one millimetre length as reported for the aquatic plant duckweed (*Wolffia globosa*) [6] to more than 110 metres height and weights of at least 2 million kilograms as reported for the redwoods¹ (*Sequoia sempervirens*). One of the oldest living plants is the King's Holly clone (*Lomatia tasmanica*) in south-western Tasmania. This plant has an extent of over 1.2 kilometres and it has been estimated that it has been cloning itself for at least 43,600 years [7]. Likewise, conifers are reported to reach an impressive age of 5,000 to 8,000 years (*Pinus longaeva*, *Taxus baccata*, *Picea abies*) [8].

¹<http://www.britannica.com/EBchecked/topic/132725/conifer>, August 2014

1. Introduction

1.1.1. Plant genome organisation

Apart from their diverse shapes, size and longevity plants also show a remarkable variation in genome organisation, like genome size, sequence composition, and chromosome number. As far as the genome size of plant species is known they range from 63 megabase pairs (Mbp) in carnivorous plants (*Genlisea margaretae*) to 149,000 Mbp in canopy plants (*Paris japonica*) [9–11]. Thus, plant genomes are about 0.02 to 48 times the size of the human genome and span a range of 2365-fold among themselves (see Table 1.1). Among them, numerous grass genomes, such as maize (2,300 Mbp), barley (5,428 Mbp), rye (8,093 Mbp), and wheat (17,100 Mbp), equal or exceed up to 5.5-fold the size of the human genome.

Common name	Scientific name	Chromosome number (2n)	Genome size (1C in Mbp)	ploidy level
<i>E. coli</i>	<i>Escherichia coli</i>	2	4.6	diploid
yeast	<i>Saccharomyces cerevisiae</i>	32	12.1	diploid
carnivorous Genlisea	<i>Genlisea margaretae</i>	52	63	diploid
thale cress	<i>Arabidopsis thaliana</i>	10	150	diploid
duckweed	<i>Spirodela polyrhiza</i>	80	158	diploid
purple false brome	<i>Brachypodium distachyon</i>	10	355	diploid
rice	<i>Oryza sativa</i>	24	489	diploid
sorghum	<i>Sorghum bicolor</i>	20	730	diploid
human	<i>Homo sapiens</i>	46	3,100	diploid
barley	<i>Hordeum vulgare</i>	14	5,428	diploid
rye	<i>Secale cereale</i>	14	8,093	diploid
wheat	<i>Triticum aestivum</i>	42	17,100	hexaploid
Norway spruce	<i>Picea abies</i>	24	19,570	diploid
canopy plant	<i>Paris japonica</i>	40	149,000	octoploid

Table 1.1.: An overview of the genome sizes and chromosome numbers of selected plant and non-plant organisms. The organisms are ordered according to their genome size in Mbp (millions of base pairs) per haploid set of chromosomes [9, 12–14].

A similar variety can be observed among related plants for the number of chromosomes shared. Chromosome number diversification is frequently encountered in flowering plants and may arise by fusions or fission of chromosomes, by accumulation of B chromosomes (see section 1.3), or by ploidy events [11]. The lowest number of chromosomes that has been reported in angiosperms, both monocots and eudicots, is $2n = 4$, while the highest chromosome numbers observed are $2n = 596$ in the palm *Voanioala gerardii* [15] and $2n = 1260$ in the fern *Ophioglossum reticulatum* [16]. In contrast, the gymnosperms have a relatively low and constant number of chromosomes ($2n = 14$ to 28), and despite their large genome size, only few polyploid species are known [17].

The chromosome number polymorphism caused by ploidy events or whole genome duplication (WGD) can be traced back to two different types: polyploidy and aneuploidy. Chromosome aberrations triggered by aneuploidy imply loss or gain of individual chromosomes, while polyploidy involve the whole chromosome complement [11]. At polyploidy level it can be distinguished between autopolyploidy, where the duplicated genome may originate from the

same species, and chromosome increasing through interspecific hybridisation with duplication of one or all involved basic sets of chromosomes (allopolyploidy). Prominent examples of economically important polyploid crops are (i) banana and apple (triploid), (ii) potato, pasta wheat, and cotton (tetraploid), (iii) wheat, oat, triticale, and kiwifruit (hexaploid), and (iv) strawberry and sugar cane (octoploid) [12].

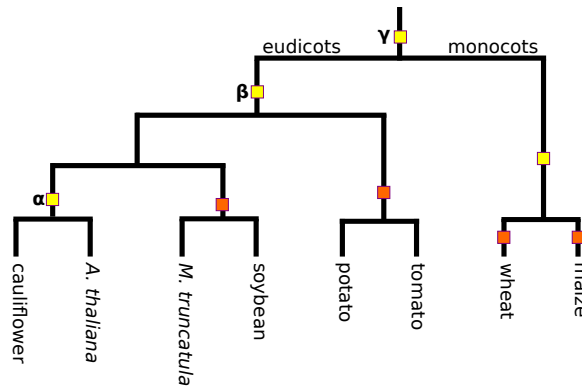


Figure 1.1.: Schematic representation of the putative locations of ancient (yellow squares) and more recently (orange squares) polyploidy events in flowering plants. The branches of the phylogenetic tree are not time scaled and the timing of the duplication events are also not specified, due to the controversial estimations in the referenced studies. The phylogenetic tree is adapted from [18–21].

Plant genomes, especially angiosperms, harbour evidences of repeated whole genome duplication events during their evolution regardless of the size of their genome [21–25]. Several studies have shown that even in the small genome of thale cress (*Arabidopsis thaliana*) at least 2 polyploidy rounds have occurred [25, 26]. Aside from various recent polyploidisation events in the maize lineage [27, 28], in soybean and *Medicago truncatula* [19, 28], in the common ancestor of the *Solanaceae* (e.g. tomato, potato) [19, 28], four major ancient whole genome duplications have been suggested, too [18, 19, 21, 22, 29]. The ancient duplication events took place (i) before the divergence of monocots and eudicots (γ event), (ii) in the core eudicots before the divergence of *Arabidopsis* from other eudicots (β event), (iii) in the ancestor of the *Brassica* and *Arabidopsis* lineages (α event), (iv) and in the common ancestor of the modern grasses (see Fig. 1.1).

Polyploidy is a major force in plant speciation and evolution, with significant implications for plants at both molecular and ecological level. The events trigger genomic instability through subsequent changes in genome structure and gene expression and can induce changes in plants' fitness and phenotype. The genomes undergo comprehensive and rapid restructuring by sequence elimination (whole chromosomes or segments) [30], sequence rearrangements [31], homoeologous and nonhomoeologous recombination [19, 22, 25], and altering DNA methylation [32–34]. The genomic reorganisations have been observed in both coding [33, 35] and repetitive regions [36, 37]. It has been suggested that transposable elements, which have been repressed in the parental lineages, can be reactivated and thereby induce sequence rearrangements with putative

1. Introduction

altering of gene expression [38, 39], generate new genes via gene duplication and amplification [39–43], or influence their transcriptional activity [44, 45]. In the coding regions, polyploidy leads to duplicated genes which may undergo various evolutionary fates, such as functional diversification, gene silencing, or preservation [1, 22, 25]. The differentiation in function implies the acquisition of a new function for one gene copy by releasing nonsynonymous substitutions under positive selection (neofunctionalisation) [46, 47] or by partition of the ancestral function (subfunctionalisation) [48–50]. Mutations within redundant genes can also lead to silencing of gene copies (pseudogenisation/nonfunctionalisation) [1, 51]. In addition, gene silencing or altered expression may be induced through insertion of transposons into the promoter region of genes [25, 52, 53]. Blanc et al. [54] have shown that the gene silencing mechanism is biased towards particular genes and gene families, like genes involved in signal transduction and transcription, and dismiss other genes, such as genes involved in DNA repair. In rare cases both gene copies and their original function are retained over long evolutionary periods [1, 22, 25]. Despite genome doubling and large gene families the number of estimated genes is comparable in plant genomes (~35,000 on average) and the genes do not contribute substantially to the increase of genome size [55–57].

The role of polyploidy events in plants evolution and speciation has been widely discussed over the past decades and open questions remain [58]. Thus far, three major advantages have mostly been associated with genome duplications. First, duplicated genes may adopt new functions that allow the altered genomes to adapt more rapidly to new environmental conditions or to expand to new ecological niches [22]. This increased genomic variability can lead to speciation under certain conditions [59]. Second, genes with an increased number of alleles may easily compensate deleterious mutations and prevent from loss of fitness [58, 60]. Third, heterosis confers benefits to allopolyploids, with characteristics that exceed that of the parental lineages [58, 61]. However, other studies disagree about the positive effects of polyploidy, arguing that genome duplication may lead to decreased fitness and adaptability due to the imposed complexity and altered epigenetic landscape [58, 59].

The increased genome size observed in plants can only partly be explained by polyploidy and segmental duplications. Repetitive elements, primarily transposable elements, have been identified as a third factor that inflate the nuclear DNA content. They make up a large fraction of the plant genomes, their amount varies widely from 10% up to 90% of the entire genome [62–64]. The transposable elements (TEs) are DNA sequences that are nested in hundreds or thousands of copies within the genome. They are able to change their positions by translocating in the genome. The TEs are divided into two classes, namely retrotransposons (class I) and DNA transposons (class II). Both classes differ in their transposition mechanism, the enzymes they require, and in the percentage of their share in plant genomes. The transposition of DNA transposons is carried by transposase enzymes, which cut them out from their location and insert them into a new target site ('cut and paste') [65]. On the contrary retrotransposons are transcribed first to RNA, and the obtained messenger RNA is reversely transcribed to complementary DNA (cDNA) and integrated back into a new location in the genome by endonuclease/integrase ('copy and paste') [65]. In plant genomes the long terminal repeat (LTR) retrotransposons outweigh other repetitive elements and are responsible for genome alteration by inducing the creation or elimination of mutations, thus activating or inactivating genes, and by increasing the genome size

through their ‘copy and paste’ mechanism [66–68]. The LTR retrotransposons can be classified in several families, which vary in their abundance among species. Most of them occur in low copy numbers, except for a few highly amplified families [69, 70]. The massive amplification of LTR retrotransposon families may cause an increase in genome size, as it has been shown in maize [70] and *Oryza australiensis* [71]. Furthermore, LTR retrotransposon families amplified in one species can be absent or present in a low number in other related species [72]. The amplification of TEs in plants and the induced genome expansion are counterbalanced by rapid DNA removal through unequal homologous recombinations and illegitimate recombination [73, 74]. It has been shown that the rate of sequence loss in plants is highly variable and that it may prevent genome ‘obesity’ [43, 67, 74].

1.1.2. Plant genome sequencing and assembly

The prior discussion already indicates that plants are organisms of amazing complexity. This complexity, as well as their distinguished significance in the world’s ecosystem, make them an extraordinarily interesting research topic. Until recently, sequencing of plant genomes used to be a time-consuming task and was only affordable for small and middle sized genomes, like *Arabidopsis thaliana* (~150 Mbp) [63] and *Oryza sativa* (~489 Mbp) [75]. The sequencing of large grass genomes (> 2 Gbp), such as barley (5.4 Gbp), wheat (17.1 Gbp) and rye (8 Gbp), has lagged behind other plant genomes so far due to their highly repetitive nature (>80%), ploidy level and complex genome organisation. The appearance of new technologies, such as next-generation sequencing (NGS, e.g. Roche 454, Illumina, SOLiDTM) and flow cytometry, has enabled the cost-efficient and rapid sequencing of large and complex plant genomes (reviewed in [8, 76–78]). The technological improvements of the sequencing platforms have increased the number of sequenced base pairs (bp) per day from thousands (Sanger technology) to millions (NGS) whilst the read length vary between 26-150 bp (Illumina) and 10-20 kbp (PacBio) [8]. Furthermore the sequencing costs have been reduced from millions to only hundred of dollars per gigabase.

Despite the new technical and economic sequencing opportunities, only few plant genomes can be considered completely sequenced to date. *Arabidopsis*, rice, maize, *Brachypodium distachyon* [79] and sorghum [80] are just some of the gold standard genomes, which have been established as model plants for other eudicots and monocots. All other recently sequenced plant genomes, like grape [81], cucumber [82], potato [83], woodland strawberry [84], tomato [85], banana [86], spruce [87], watermelon [88], and chickpea [89], can be regarded as drafts at various stages of sequence completion. Although sequence sequencing of large plant genomes is no longer a problem, storage capacity and assembling of the massive data generated by NGS remains a challenge that needs to be overcome to achieve a complete and accurate genome assembly [77].

The high repetitivity [67] and ploidy level [90] of plant genomes make assemblies difficult, since the nearly identical sequences can often collapse on top of one another, even if they are located at different sites in the genome. Furthermore, plant genomes tend to have a complex gene content with large gene families and pseudogenes, which are nested between various-sized

1. Introduction

blocks of repetitive elements [8]. Thus, an accurate assembly is difficult to achieve and often restricted to the low-copy regions, which results in highly fragmented draft genome assemblies [76, 91]. The assembly quality may be negatively affected by DNA contamination and sequencing errors, too. The raw sequence data is often contaminated with vector/adaptor sequences, organellar DNA, or human DNA, hampering a successful assembly and requiring a pre-processing step to remove them [8, 77]. The impact of sequencing errors can be compensated by a high sequencing coverage (80 to 100-fold), shifting the bottleneck toward computational speed and memory usage.

In spite of the numerous assembly algorithms (e.g. ALLPATHS-LG [92], SOAPdenovo [93], Abyss [94]) none of them can deal with all mentioned impediments yet [77, 95, 96]. Still, the generated draft genomes represent a valuable resource for genetic and genomic applications and can be used to provide insight into their gene and repeat content. In addition, through interspecific comparisons the structure of their genomes, as well as their evolutionary processes and phylogenetic patterns can be revealed [77, 91]. They also support the design of molecular markers, but may fail to address accurate SNP calls [91, 97].

In a first attempt to tackle large grass genomes, like barley [98, 99], wheat [100, 101] and rye [102], low-coverage (~1 to 5-fold) chromosomal/genome survey sequences (CSS/GSS) have been generated using the Roche 454 sequencing platform. Because of the low sequence coverage the construction of an accurate assembly was not feasible. Therefore, there was an urgent need to design and implement new methods to assess the gene content and to position the identified genes along the individual chromosomes. In this thesis a novel, powerful approach will be presented (see Chapter 2), that makes use of closely related model grass genomes to deduce the gene order in the genomes under investigation and to provide a subassembly of the gene-rich regions of the genomes ('gene-ome').

1.1.3. Taxonomy, origin, and economic importance of grasses

The grass family (*Poaceae*) is one of the most economically important and ecologically dominant families of monocotyledonous flowering plants (angiosperms). The members of this family are widespread throughout the whole world, covering at least 20% of the earth's surface along different climate zones [103]. The appearance of grass pollen in the fossil records has allowed to trace their origin back to roughly 50-77 million years ago (mya) [104, 105]. These results have been confirmed by sequence comparisons of chloroplast and ribosomal DNA between several grass species [106, 107]. Their edible grains and prevalence have given cause to their domestication and cultivation over the past thousands of years establishing them as a source of staple food and feedstock for livestock. Recently, grasses have been taken into consideration as renewable and ecofriendly alternatives to fossil fuel, too.

The monocot *Poaceae* family comprises over 10,000 species grouped in ~650 genres [104] that are classified in several subfamilies. Among them are the *Pooideae*, *Ehrhartoideae*, and the *Panicoideae* subfamilies. These three subfamilies include all the major cereals (see Fig. 1.2) and can be subdivided in several tribes, like (i) *Triticeae* (barley, wheat, rye), *Aveneae* (oat), *Brachypodieae* (bromus) and *Poeae* (ryegrass, fescue) (*Pooideae* tribes), (ii) *Oryzaceae* (rice) (*Ehrhartoideae* tribe), and (iii) *Andropogoneae* (corn, sorghum, sugarcane) and *Paniceae* (pearl millet,

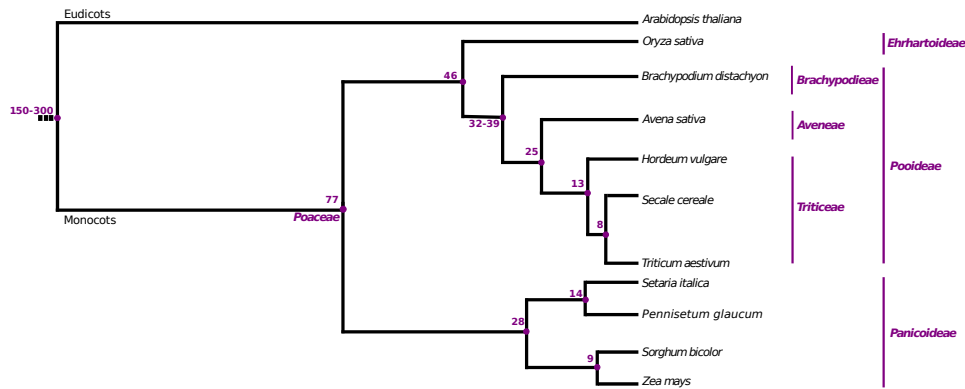


Figure 1.2.: Schematic representation of the phylogeny and divergence time of several well-studied and agronomical important members of the *Poaceae* family (figure adapted from Gaut et al. [105]).

foxtail millet) (*Panicoideae* tribes). These crop plants are extensively grown and supply more than 60%² of the calories ingested by the world's population [108]. Corn holds the leading position with an annual production of about 1000 million tons, followed by rice with 745 million tons and wheat with 713 million tons² (see Fig. 1.3).

Maize (*Zea mays*, $n = 10$ chromosomes) is one of the most versatile crops which is processed not only to oil, syrup, flour, starch, and forage but also to biogas and fuel-ethanol. The crop grows all over the world in different climate zones, such as tropical, subtropical, and temperate climates. Maize has a monophyletic origin and descends from an ancestral teosinte which has most probably been harboured in the Balsas river drainage in southern Mexico [109]. Genetic and archaeological data support this origin of maize domestication. The beginning of its domestication is estimated for the period between 6,000 and 10,000 calibrated years before present (cal BP) [109–111].

Rice (*Oryza sativa*, $n = 12$ chromosomes) is the only grain crop that is cultivated solely for use as human food. Its consumption supplies the need of roughly half of the world's population [112]. There are two major subspecies of cultivated rice: the long, thin grained *indica* (*Oryza sativa* ssp. *indica*) and the short, thick grained *japonica* (*Oryza sativa* ssp. *japonica*) [113]. The subspecies *japonica* adapts to cool climates in northern China and south-eastern Asia, while, in contrast, the subspecies *indica* prospers in hot, tropical climates [113]. The cultivation of rice was initiated between 8,000 and 10,000 years ago, followed by a gradual domestication in eastern China and northern India [114–117].

In many cultures, wheat substantially contributes to the daily diet, supplying at least 20% [118] of the consumed calories. It is rich in starch (60-80%), proteins (8-15%), minerals and vitamins, although it lacks four essential nutrients: vitamins A, B12, C, and iodine [113]. Several wheat types are cultivated world-wide, such as Einkorn wheat (*Triticum monococcum*), Timopheev's wheat (*Triticum timopheevii*), *Triticum turgidum* with the varieties Rivet wheat

²<http://faostat.fao.org/>, September 2014

1. Introduction

(*Triticum turgidum* L. ssp. *turgidum*), Emmer wheat (*Triticum turgidum* L. ssp. *dicoccum*) and durum wheat (*Triticum turgidum* L. ssp. *durum*, ‘pasta wheat’), and *Triticum aestivum* with the varieties Spelt wheat (*Triticum aestivum* L. ssp. *spelta*) and bread wheat (*Triticum aestivum* L. ssp. *aestivum*). Here, bread wheat makes up the largest contingent with over 93% [118] of the wheat area under cultivation. Wheat is an adaptable cereal crop which is extensively cultivated across different climatic zones, i.e. from the equator to near Arctic regions. However, it grows best in temperate areas and the main wheat producers are consequently located in such temperate regions (Mediterranean basin, southern Russia, central United States, Argentina, south-western Australia). The origin of wheat lies, according to archaeological findings and DNA fingerprinting, in the Neolithic south-eastern Turkey and northern Syria ~10,000 years ago [119–122]. Einkorn wheat and Emmer wheat are the oldest known cultivated wheat species, while other types of wheat followed later with advances in agriculture.

On the fourth rank in cereal production there is barley (*Hordeum vulgare*, n = 7 chromosomes) with an annual production of 145 million tons³. The largest proportion of barley harvest (75%) is consumed as forage, followed by its use in alcoholic and non-alcoholic beverages (20%) and as an ingredient of food products (5%). The high content of soluble dietary fibre in its grains has classified barley as a functional food that can be efficiently used to improve health and to reduce the risk of diet-related diseases such as diabetes, cardiovascular complaint, and cancer [123]. Barley is a diploid, selfing species with a genome size of 5.1 gigabases (Gbp) [124]. Its ability to grow in varying geographic regions, its tolerance to salinity, drought, and cold, as well as its close relationship to wheat are the reasons why it is considered a model of plant genetic research [125]. Furthermore, barley is one of the oldest known cultivated cereal grains and has its origin most likely in the Fertile Crescent of southwest Asia (Iraq, Turkey, Iran, and Syria) [119, 126]. Badr et al. [127, 128] proposed that the cultivated barley has a monophyletic origin and suggested that the involved regions outside the Fertile Crescent are rather centres of barley diversification than centres of domestication. However, this hypothesis is contradicted by several studies, which indicate that barley does not have a single centre of origin [129–132]. Its domestication is dated back to the period from ~8,500 to 13,000 years ago based on archaeological evidences of human sites and on evidences of living wild progenitors [113, 119, 120, 133–135]. The wild species *Hordeum spontaneum* is considered the progenitor of domesticated barley [119]. Both species show close genetic affinities and differ mainly in their modes of seed dispersal: the two-rowed, hulled wild type shatters seeds, whereas the six-rowed, naked cultivar does not [113, 136].

Rye (*Secale cereale* L.), another member of the tribe *Triticeae*, is not as widespread as wheat and barley. Still, it is about as successful as barley in northern and eastern Europe. There, it is mainly used as fodder and for bread and malt production (which is a key ingredient for the manufacture of whiskey and beer). Rye is a hardy plant with an elevated tolerance to frost, drought, and poor quality soils, such as acid and/or sandy soils [113, 137]. Besides, it comprises annual and perennial cultivated, wild, and weedy species, that show cross-pollinating and self-pollinating features. The classification of the species within this genus is not completely resolved and relies mainly on morphological characteristics. Hence, depending on the used species definition, the genus *Secale* can be divided up to 15 species [137–140]. However, the

³<http://faostat.fao.org/>, September 2014

most widely accepted ranking of the rye taxa comprehends four taxonomic groups, namely the annual outbreeder *Secale cereale* L., the annual inbreeder *Secale vavilovii* Grossh. and *Secale sylvestre* Host, and the perennial cross-pollinating *Secale strictum* (Presl) Presl (syn. *montanum*) [137, 140]. This classification is in agreement with the result of several studies based on the analysis of ITS rDNA sequences [141], microsatellite markers [142], AFLP markers [143], RFLP of the mitochondrial DNA [144], and ISSR and IRAP markers [145]. Nevertheless, due to the high similarity established between *Secale cereale* L. and *Secale vavilovii* Grossh. scientists tend to classify the last-mentioned as a subspecies of *S. cereale* [139, 143]. The known rye cultivars belong only to one out of the four listed taxa, namely to the species *Secale cereale* ssp. *cereale*. Similar to other Triticeae cereals, the Fertile Crescent (especially the area of eastern Turkey) have been considered the cradle of rye development and cultivation, too. Growing as a weed in barley and wheat fields, the unintentional cultivation and domestication of rye started around 7,000 to 10,000 years ago [134, 137, 146, 147]. The gradual transition from weed to cultivated crop has been favoured by its ability to grow in arid, stony, and even sandy soils, its frost-tolerance, and its fully shattering grains [134]. Even today, in Anatolia, Syria, Iran and Iraq wild rye forms infest wheat fields, enabling introgression into cultivated crops. According to remains of rye grains at archaeological sites, the rye migration to Europe most probably had taken place in several stages five to eight centuries before present as a weed among other cereals [137, 147, 148].

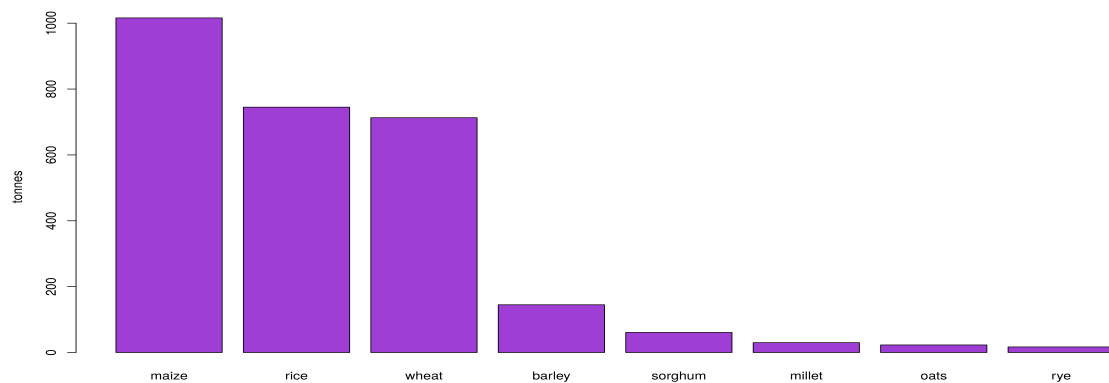


Figure 1.3.: The worldwide production of selected crop plants for the year 2013 (Food and Agriculture Organisation of the United Nations, <http://faostat3.fao.org/>).

As already mentioned above, cereal crops constitute an indispensable source of food, providing calories and essential nutrients for the world's diet. Starting with their cultivation and domestication thousands of years ago, cereals have been the target of several attempts to adapt them to new climatic conditions and to increase their yield. The latest successful attempt was about 50 years ago in the so-called 'green revolution'. The use of new varieties, better irrigation systems, fertilizer, and pesticides led to a substantial harvest increase, but to some extent these improvements were at the expense of the environment [149]. Nowadays, breeders and researchers have to face new challenges to compensate previous mistakes. They need to halt the decline in crop yields and, if possible, increase it again. Besides a rapidly expanding population, climatic changes (higher temperatures, pollution), the degradation of arable land (increased soil

1. Introduction

acidity, salinity, and heavy metal), and the higher susceptibility of crops to diseases and pests impede their task to secure food production [149]. Thus, the understanding of the genome evolution, structure, and organisation of cultivated and wild crop plants will enable the development of new cultivars with an improved genetic variation, disease resistance, and productivity.

1.1.4. Comparative genomics in grasses

The comparisons between two or more organisms have improved over the past decades with the advent of genetic markers (e.g. RFLP and microsatellite marker) and DNA sequences (e.g. ESTs: expressed sequence tags, BAC: bacterial artificial chromosomes). They enabled to move from simple trait observations to genome structure and function comparisons. Thereby, the terms conserved synteny (Greek: syn = together, taenia = ribbon) and collinearity have been used to characterise orthologous genes located on related chromosome regions within two species (synteny) and which are conserved in the same order (collinearity). The grass family (*Poaceae*) has served as a 'model system' for comparative studies with the aim to understand the taxonomic relationships, evolution, speciation, structure and organisation of plant genomes [150, 151]. Furthermore, the comparative data has been used to develop DNA markers for marker-assisted breeding and to transfer the knowledge about agronomic important trait genes (e.g. disease resistance genes, genes for yield and nutritional enhancement) from a model organism to other species [152, 153].

Despite tremendous diversity in chromosome number (from $n=5$ to $n=21$), ploidy level (from $2x$ to $10x$) and genome size (from 355 Mbp to 17,100 Mbp), comparative analyses of grass genomes unveiled a striking collinearity among the *Poaceae* [150, 154–158]. Grass genomes, like that of wheat, barley, rye, sugar cane, rice, maize, sorghum, and oat, can be dissected in syntenically conserved chromosomal fragments (linkage blocks) and these segments reassembled to any grass genome by simple rearrangements [154, 156]. The linkage blocks may allow to determine lineage-specific evolutionary events (e.g. gene duplication, losses, and introgression) and to deduce the karyotype of the last common ancestor [159]. The high collinearity observed between the orthologous linkage blocks is often disturbed by small rearrangements, such as deletions, translocations, insertions, and duplications. These small rearrangements can reshuffle genes and impede the transfer of knowledge from model genomes (e.g. rice) to other species of interest.

Comparative analyses induced by low-resolution marker maps have indicated several lineage-specific chromosomal rearrangements in grasses. Chromosome reshuffling has been observed in both outbreeding and inbreeding species, like rye [102, 160, 161], wheat [100, 162], *Aegilops umbellulata* [163], *Aegilops longissima* [164], *Triticum timopheevii* [165], *Aegilops tauschii* [166], ryegrass (*Lolium perenne*) [167, 168], meadow fescue (*Festuca pratensis*) [169, 170], and oats [171, 172] (see Fig. 1.4). The number of alterations among species is uneven (from one to eleven) and does not correlate with the degree of relatedness [151]. Although it is unknown why certain species accumulate structural changes more frequently than others, the high degree of conservation within these rearranged segments is retained [155].

For example, the genomes of rye and *Ae. umbellulata* have undergone extensive rearrangements (e.g. reciprocal translocations, inversions) relative to wheat involving all chromosomes,

except chromosome 1 in rye [161, 163]. Following these rearrangements Devos et al. [161] have proposed a model of rye evolution. According to that model a first ancient 4L/5L translocation took place in various *Triticeae* species, followed by a second step of 4RL/7RS and 3RL/6RL translocations. The translocation order of a last event remain unresolved, due to various possible scenarios: (i) the 2RS/6RS translocation was followed by two more translocations (6S/7L and 2S/4L) and a pericentric inversion of chromosome 6R, or (ii) the 2RS/7RL and 6RS/4RL translocations took place first followed up by the 2RS/6RS translocation [161]. In *Ae. umbellulata* each individual chromosome revealed conserved synteny with up to five wheat chromosomes [163]. The chromosomes 1U and 2U are collinear with almost all of the wheat chromosomes 1 and 2, while for the short arm of chromosome 3U a translocation on chromosome 7U has been suggested. Chromosome 4U is homoeologous to the wheat chromosomes 7DL, 4DL, and 6D. The chromosomes 5U and 7U show collinearity to the wheat chromosomes 4DL and 5D, and 3DS and 7D, respectively. Most of the rearrangements seem to occur on chromosome 6U, which shares homoeology with the long arm of wheat chromosomes 1DL, 2DL, 7DL and partially with both chromosome arms of 4D and 6D [163]. Anyway, grass species with large number of rearrangements are rather an exception than the norm.

In ryegrass and meadow fescue comparisons with wheat and barley have revealed an unidirectional translocation that relocates a segment from the long arm of chromosome 5 at the distal end of the short arm of chromosome 4 (4S/5L) [167–170]. Kopecky et al. [170] hypothesised that this translocation occurred in the *Triticeae* lineage after the split from the *Festuca* and *Lolium* genera. Chromosomal rearrangements can be observed in wheat, too. Chromosome 4A underwent several rearrangements, involving two translocations with the chromosome arms 5AL and 7BS, as well as a pericentric inversion [100, 162].

DNA sequence-based comparisons equally revealed insights into the grass gene organisation. According to several studies, the grass genes are not randomly distributed along the chromosomes and the size of the intergenic regions varies in accordance with their genome size [57, 174–179]. Most of the genes are clustered into gene islands of fluctuating density and size [57, 179, 180]. The gene-clusters can alternate with single genes and they show a greater density at the distal part of the chromosome arms by having shorter ‘inter-insular distances’ [178]. For example, genome comparisons of sorghum and *Aegilops tauschii* have revealed that the islands contain 3.7 to 3.9 genes on average and a mean distance between the islands of 15.1 to 205 kilobase (kb) [178]. In wheat, Raats et al. [180] have shown that the number of genes increase in the telomeric regions up to 5.9 genes/Mbp, while in the centromere area only an average of 2.5 genes/Mbp can be observed. The distribution of genes at the distal regions of the chromosome arms have been confirmed in several sequenced grass genomes, like rice [181], sorghum [80], maize [182], *Brachypodium distachyon* [79], and *Setaria* [183]. The size of intergenic regions between two genes deviate between small and large grass genomes (see Fig. 1.5). In the large genomes of *Aegilops tauschii* (4.9 Gbp) and maize (2.6 Gbp), the gap between the genes is, with 16.5 and 140 kilobase (kb) on average, several times bigger than in the smaller genomes of sorghum (700 Mbp) and rice (489 Mbp) [12, 174, 178]. These regions mostly consist of transposable elements (TEs), which are mainly responsible for the observed genome size variation. While small grasses have a low abundance of TEs (e.g. ~40% in rice and *Brachypodium distachyon*, ~60% in sorghum), large genomes (>2 Gbp) can exhibit an abundance of over 80%

1. Introduction

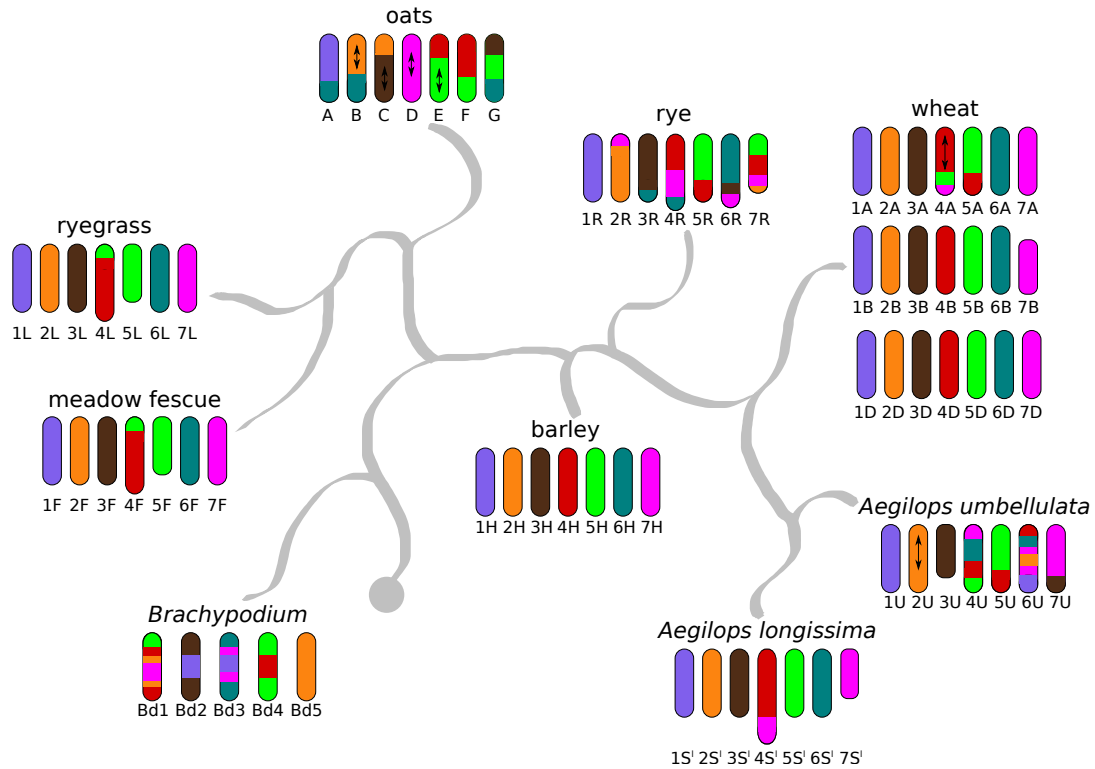


Figure 1.4.: Schematic representation of genome reorganisations in several grass genomes depicted in phylogenetic context [79, 161–164, 168, 169, 171–173]. The rearrangements (translocations and inversions) of the syntetically conserved linkage groups are depicted relative to that of the barley genome. Inversion of segments are indicated by double-headed arrows.

TEs, as it has been shown for maize and wheat [79, 80, 176, 182, 184].

1.2. The *GenomeZipper* approach

1.2.1. Motivation

Plant genomes, especially cereals, are complex and challenging to study due to their large and often polyploid genomes, the huge amount of repetitive sequences, and costs. A first milestone to tackle the complex structure of crop plants, has been achieved by using the next generation sequencing (NGS) technologies (Roche 454, Illumina) and flow cytometry. The massive amount of generated chromosomal survey sequences and the highly repetitive sequence content pose new challenges in obtaining an accurate assembly and analysing the low-complexity region of the genome. The aim of this thesis is to develop a virtual approximation of gene order along the chromosomes of complex cereal genomes. This is achieved by the identification of highly

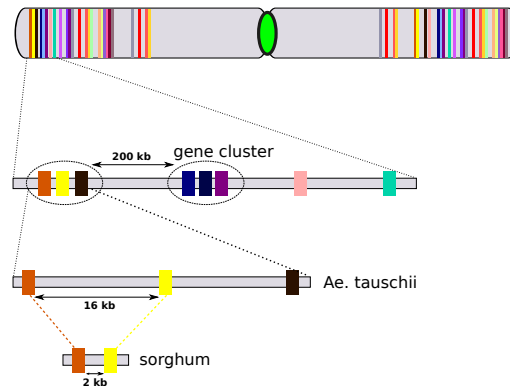


Figure 1.5.: Figure illustrates the organisation of conserved coding regions in grasses. The distal areas of a grass chromosome show a more dense gene distribution than its pericentromeric and centromeric regions. The genes are organised in gene-rich islands and single genes which are interspersed by large intergenic regions (e.g. 2 to 16 kb in sorghum and *Ae. tauschii*). The size of the intergenic regions is correlated with the genome size, so that the distances between two genes or gene islands are shorter in small genomes than in large genomes (figure adapted from Feuillet et al. [177], [57, 174, 178]).

conserved genic sequences based on homology to closely related reference genomes. The approximated order has been inferred from dense genetic marker maps of the species itself and for regions without marker support from the orientation of the corresponding homologous hits in the reference genomes.

1.2.2. The concept

The annotation and exploration of large plant genomes, such as barley, rye, and wheat, has lagged behind those of other organisms due to their striking complexity as well as technological and economic restrictions (see sections 1.1.1 and 1.1.2). With the remarkable progress of new technologies, like NGS and flow cytometry, their genomic sequences became accessible and new insights into the structure of complex genomes could be gained. The focus of this thesis is on the genomes of barley and rye and the development of a novel approach to overcome the impediments caused by the lack of physical maps and incomplete assemblies.

The barley and rye chromosomes have been shotgun sequenced using Roche 454 GS FLX sequencer at an average sequence coverage of 1.3-fold and 1.04-fold, respectively [98, 99, 102]. The low-coverage and high repetitivity of the millions of read sequences, as well as the lack of physical maps posed a challenge to process, accurate assemble, and annotate the two cereal crop genomes. The identification of low-complexity sequences among the massive amount of chromosomal survey sequences was rather straight-forward and could be solved using well-known standard methods, like sequence similarity search. Nonetheless, the position of these

1. Introduction

gene fragments along the individual barley and rye chromosomes remains mostly unknown, except few sequence reads that have been associated with genetic SNP markers.

The novel approach called *GenomeZipper* addresses the lack of order and structure among the identified low-complexity sequences by reverse engineering. The algorithm makes use of the high degree of homologous genes among grasses that are located on corresponding chromosomes (syteny) and that are in conserved order (collinearity) (see section 1.1.4). The novelty of this method lies in the expansion of genetic marker maps with positional information from syntenic collinear blocks of fully sequenced reference genomes and in the positioning of chromosome-specific shotgun sequences along the new built virtual chromosome structure. The exploitation of syteny and collinearity information increases the number of ordered gene loci from a few to several hundreds in a simple and effective manner. The inferred individual gene loci are in addition associated with supplementary genomic information, such as full-length cDNAs and ESTs. Thus, the *GenomeZipper* successfully interlinks heterogeneous data sets to establish a robust approximation of the gene positions and order for species without assembled genomes.

The *GenomeZipper* approach has been applied on each individual rye and barley chromosome/chromosome arm. The fully sequenced genomes of *Brachypodium distachyon* [79], rice [75], and sorghum [80] have been chosen as reference genomes. The order of the syntenically conserved orthologous genes in the virtual gene map of rye and barley took in consideration the evolutionary relationships between the regarded genomes. The collinear genes of the evolutionary closest reference genome, namely *Brachypodium*, got the highest rank, followed by rice and sorghum genes. After the integration of the syteny information as described above, additional evidences, such as barley full-length cDNAs [185] and rye and barley EST collections (<http://www.harvest-web.org>), were attached to the ordered gene scaffold. At last all rye and barley 454 sequence reads with an homology match to the markers, collinear genes, full-length cDNAs, and ESTs were selected and positioned along the virtual chromosomal structure in a stringent (best bidirectional hit) and less stringent (first best hit) manner.

The generated rye [102] and barley [98, 99] gene maps provide a high quality surrogate for working genomes and closely related grasses until their physical maps and complete genome sequences become available. They have enabled comparative genomics, leading to new insights regarding the chromosome structure and organisation, as well as evolutionary relationships and dynamics among grasses. Furthermore, the linear ordered gene maps provide a valuable resource to plant researchers and breeders for marker development, chromosome dissection, and physical map anchoring.

1.3. B chromosomes

1.3.1. Motivation

Despite being intensively researched over the past century, B chromosomes still remain an evolutionary mystery and raise questions concerning their mechanism of inheritance, effects on host,

molecular composition, and origin. Recent technical advances, like chromosome flow sorting and next generation sequencing, enable the first low coverage sequencing of a B chromosome, namely the rye B chromosome. Since the rye B chromosome lacks a complete assembly as well as genetic and physical maps its analysis remains constrained.

This thesis attempts to make use of the rye and barley *GenomeZipper* outputs to overcome these limitations and to provide a comprehensive understanding of the rye B chromosome origin and sequence composition and organisation. The rye and barley virtual ordered gene maps are used as references to apply detailed sequence comparisons against the rye B chromosome. These comparisons revealed the DNA components and ‘founder’ A chromosomes of the rye B chromosome. Additionally, a comprehensive model of B chromosome evolution is proposed.

In the next sections, a detailed introduction to B chromosomes is given. The outcome of the comparative analysis mentioned above is described and discussed in section 3.3 and 4.2.2.

1.3.2. Definition of B chromosomes

Supernumerary accessory chromosomes, called B chromosomes (Bs), are optional extras to the basic A chromosome complement of a cell. Since their first discovery in insects by E. B. Wilson in 1907 [186] the number of described B-carrying species has been increasing continuously, as well as the varying terminology used for them. Denominations like accessory chromosomes, supernumerary chromosomes, additional chromosomes, *k*-chromosomes, extra-chromosomes, had been used until the term B chromosomes proposed by Randolph [187] was accepted [188]. Since then various aspects of their biology have been well documented and catalogued, of which the most comprehensive atlases were realised by Battaglia [188] in 1964 and Jones & Rees in 1982 [189]. They differ morphologically from the standard A chromosomes and are dispensable for the normal growth and reproduction of the organism carrying them [189]. B chromosomes can be present or absent among individuals of the same population (e.g. *Secale cereale* $2n = 14 A_s + 0-8 B_s$; *Vulpes vulpes*, $2n = 34 A_s + 0-8 B_s$; *Rattus rattus*, $2n = 42 + 0-5 B_s$). To prevent the Bs from being eliminated from a population, they behave like parasitic elements ensuring their transmission and survival by a selfish ‘drive’. In most cases their effect on the host is triggered by the number of Bs present within the organism. In a low number, little or no impact on the host can be observed, while exceeding a species-specific number the phenotype and fertility are negatively influenced [189]. Furthermore odd or even-numbered B chromosome combinations influence the phenotype to varying degrees (see section 1.3.4). Thus, the B chromosomes are an extraordinary example of a numerical chromosome variation with distinct features, like scattered distribution within species and failing to pair with A chromosomes at meiosis, which distinguishes them from another chromosome number polymorphisms, like aneuploidy ([189], reviewed in [190–193]).

Although they lack obvious adaptive properties, B chromosomes are widespread over all eukaryotic groups. Nonetheless, Bs may be easily overlooked in species with limited karyotypic information and hence their occurrence may be underestimated. Further development of technologies will facilitate more detailed analysis and discovery of new B-containing species. Additionally, the absence in most individuals of a population or even in certain organs of a species

1. Introduction

impedes this estimation, too. A telling example for this is *Aegilops speltoides* which carry Bs only in aerial organs while the roots have none [194].

In plants the first B chromosomes were described in the 1920s, namely in rye (1924) [195] and maize (1925-1928, [187, 196, 197]). Later, B chromosomes were found to occur in several groups of gymnosperms and angiosperms, showing a large disparity between their presence in monocots (8%) and dicots (3%). The families with the largest number of B chromosomes are the *Poaceae* and the *Asteraceae*, with hotspots in grasses and lilies [191]. They have been encountered in outbreeding but not in inbreeding species and cultivars [189, 198]. The findings show an equal distribution across diploids and polyploids and a strong correlation between genome size and B frequency in plants. For the absence of Bs in smaller genomes, two explanations were given: (i) species with small genomes cannot cope with the impact of Bs, and (ii) large genomes with their high amount of non-coding DNA provide a better source for the creation and evolution of Bs [198, 199]. Within the mammals, Bs are more frequent in species with acrocentric chromosomes [200], while in humans no B chromosomes have been discovered so far. However, some of the human structurally abnormal chromosomes, so-called small supernumerary marker chromosomes (sSMC), show B chromosome-like traits and have the potential to become Bs [201].

Most B chromosomes are smaller than the A chromosomes. In a few cases the size of Bs vary from tiny (microchromosomes: daisy *Brachycome dichromosomatica* [202]) to same size like the corresponding A chromosomes (*Sorghum nitidum*, roach *Rutilus rutilus*), and only on rare occasions it exceeds the largest A chromosome (cyprinid fish *Alburnus alburnus* [203]). The maximal number of B chromosomes tolerated in natural populations varies widely among the species. The changes observed in B chromosome frequencies between species depend on several factors, like environmental conditions and accumulation mechanisms [204, 205]. In wild species, such as the grass *Lolium perenne* [189], *Brachycome dichromosomatica* [206], flatworm *Polycelis nigra* [207] and grasshopper *Eyprepocnemis plorans* [208], rarely individuals sustaining more than three Bs have been encountered. Up to 34 B chromosomes have been found in corn plants (*Zea mays*), followed by 20 Bs in chives (*Allium schoenoprasum*), and 15 Bs in the sea campion *Silene maritima* (see [189] for references) and the frog *Leiopelma hochstetteri* [209].

1.3.3. Origin of B chromosomes

Despite their discovery over a century ago B chromosomes remain a mystery. The knowledge of the origin, evolution, composition, regulation, and accumulation, as well as the role they play in host species is still limited and unanswered for the Bs of most species. Technological advances in DNA sequencing during the last decades allowed to address these matters and provided valuable insights into the B chromosomes, even if further studies are necessary in order to completely understand them.

Regarding their origin, it is widely accepted that B chromosomes arise most likely from the standard complement of A chromosomes [188–191, 204], although they do not have a single, common mode of origin in all species. During the last decades, several plausible scenarios

and hypotheses of B chromosome formation and evolution as autonomous components of the genome were proposed [209–213]. These hypotheses postulate that B chromosomes can be derived from autosomal [210, 214–216] or sex chromosomes [209, 217, 218] promoted by intraspecific variation, or as a spontaneous by-product of interspecific hybridisation [212, 213, 219]. All scenarios presume that DNA fragments escaped from selective constraints develop toward B chromosomes [213]. Similar chromosome Giemsa-banding patterns and comparisons of repetitive DNA sequences in B and A chromosomes, phylogenetic analysis of transposable elements, as well as experimental data provide evidence and support for these theories.

The most studied supernumerary chromosomes support an intraspecific origin involving autosomes. In order to become B chromosomes, the A chromosomes need to undergo the stage of polysomy (Hare's-foot Plantain [211], garden pea [214]), the formation of centric fragments by chromosomal rearrangements (whiptail catfish [220]), or be induced by an amalgamation of several A chromosomes (*Brachycome dichromosomatica* [221], maize [215]). One of the first fully documented mode of B chromosome origin has been described in an experimental population of Hare's-foot Plantain (*Plantago lagopus*) by Dhar et al. [211]. The B chromosome resulted from a rapidly altering trisomic chromosome, that during the formation accumulate telomeric repeats, massive amplified 5S rDNA sequences, and a functional centromere. All these accumulations confer stability and isolation from recombination with the regular A chromosomes at meiosis for the new formed chromosome. The heterochromatic structure of the isochromosome also indicates genetic inactivity and lacks any visible effects on the plant phenotype [211]. A similar model of B chromosome formation from trisomics or trisomic fragments has been equally described in the garden pea (*Pisum sativum L.*). As a source of B nascent Berdnikov et al. [214] proposed an extra chromosome of tertiary trisomic composed of the short arms of chromosomes 3 and 6. These additional chromosomes undergo molecular degeneration, like the shortening of telomeric regions, accumulation of repetitive elements, deletions and mutations in genic regions, and the loss of their ability to pair with homologous chromosomes, in order to turn into a true B chromosome [214].

Additionally to autosomes, sex chromosomes can also provide a source for the genesis of B chromosomes as it has been proposed in the relict frog *Leiopelma hochstetteri* [209, 222], in Cichlid fishes [223], and in the grasshopper *Eyprepocnemis plorans* [217]. The arise of Bs from sex chromosomes has been assumed due to the ability of Y-chromosomes to be absent from the genome and a similar evolutionary mechanism for both chromosomes has been hypothesised. Furthermore, cytogenetic and DNA composition analyses of B chromosomes support the coincidence between Bs and sex chromosomes. In the grasshopper *E. plorans* the order and location of B-located DNA regions show a high similarity to its sex chromosomes [217]. The New Zealand frog *Leiopelma hochstetteri* gained its B chromosomes most probably from the devolution of the univalent W sex chromosome, due to its high variation in size, morphology and heterochromatin distribution [209, 222]. The presence of Bs can also influence sex determination generating female-biased sex ratios as shown in cichlid fishes [223].

Another theory of B chromosomes origin presumes that they evolved from foreign DNA provided from closely related species through interspecific hybridisation. It is assumed that hybridisation induce structural rearrangements which supply the required sequences to form a B chromosome [190]. There are several plausible mechanisms for the formation of B chromo-

1. Introduction

somes. Bs can emerge from introgression and/or allopolyploidisation events (i) as fragmented components of the donor regular chromosome set, (ii) by direct transfer from the donor to the new species, or (iii) as a side-product of chromosomal rearrangements and recombinations between host and donor [213]. Though all these scenarios are conceivable, in most cases it remains unclear how the transition to Bs occurs and its verification turns out to be difficult. Interspecific hybridisation as B origin has been suggested in the parasitic wasp *Nasonia vitripennis* by McAllister et al. [213]. By using phylogenetic analysis the authors show that the supernumerary paternal-sex-ratio (PSR) chromosome has recently been transferred to *Nasonia vitripennis* from another *Trichomalopsis* species. Later, Perfectti et al. [219] provided experimental evidence for hybrid origin of PSR-type Bs in jewel wasp, too. They show that the introgressive transfer of a chromosome region from *Nasonia giraulti* into *Nasonia vitripennis* and its subsequent fragmentation results in the creation of a new B chromosome with phenotypic effects. In jobs tears (*Coix lacryma-jobi*) Sapre et al. [212] concluded that B chromosomes could originate from spontaneous hybridisation. They show that homologous and/or non-homologous chromosomes can flow between related species and behave like B chromosomes in the new host. Another possible example is the Amazon Molly fish *Poecilia formosa*, an all-female gynoform that depends on sperm of males from related species to trigger embryogenesis. The observed microchromosomes appear to incorporate foreign subgenomic DNA from the sexual host species to compensate the disadvantages of asexuality, like the accumulation of deleterious mutations [224].

Species in which the Bs may be traced back to certain donor chromosomes are rare. It is also unclear if different types of B chromosomes have the identical provenience in the same species or not. We can presume that the formation of a new B chromosome interacts with the evolution of karyotypes, as observed correlations between B nascency and large genomes with low chromosome number [198, 199] suggest. To gain an accurate and detailed picture of Bs formation further investigations are necessary.

1.3.4. B chromosome effects

In many species B chromosomes show less or no obvious effect on the phenotypes. Effects associated with phenotypes of their hosts are rare and difficult to identify, due to variations resulting from different environmental conditions and genetic heterogeneity of the A chromosomes [189]. Over the years, however, a number of studies have reported on species that act on B chromosomes, although the observed effects cannot be clearly attributed either to their presence, number or gene activity. An increased number of Bs may affect the fitness of the host, like growth, vigour, germination, and fertility, as a result of energy cost and interference with As during meiosis [189].

The influence of Bs range from neutral to harmful, depending upon the number present in the carriers and their odd or even-numbered combinations. In a high number their effects contribute detrimental to fertility and generate distinguishable abnormal phenotypes. Individuals that bear odd-numbered Bs are often more adversely affected than those with even-numbered irrespective of their numbers. Species responsive to the presence of Bs were found in flowering plants (chives, *Haplopappus gracilis*, *Plantago coronopus*), grasses (*Aegilops speltoides*, *Avena sativa*,

Aegilops variabilis, *Secale cereale*, *Zea mays*), animals (*Leiopelma hochstetteri*, cichlid fishes, *Polycelis nigra*) and fungi (*Nectria haematococca*) (for references, see Table 4.2 of [189], [193]). In *Haplopappus gracilis* [225] Bs change the pigmentation of the achenes from brownish red to dark purple, while in *Plantago coronopus* one B chromosome induces complete male sterility. B chromosomes reduce growth, vigour and fertility of the host of a number of species, like in rye, maize [226], and *Ae. speltoides* [194]. They control crown rust resistance in *Avena sativa* [227] and homoeologous pairing in hybrids between common wheat and *Ae. variabilis* [228]. In maize, an increasing number of Bs are responsible for white leaf stripes and narrow leaves [229]. In animals, the presence of Bs affect cocoon production and juvenile growth in *Polycelis nigra* [230], reduce the larval size in black flies *Cnephia ornithophilia* [231], and play a role in sex determination in the cichlid fishes [223] and the Hochstetter's frog (*Leiopelma hochstetteri*) [218]. Recently, another negative effect of Bs on hosts has been described in *Acanthophyllum laxiusculum*, where Bs reduce the antioxidative response of the plant to salt stress [232].

Due to the mostly negative influence on species, Bs are considered parasitic regardless their number, except for chives (*Allium schoenoprasum*), ryegrass and a fungal pathogen (*Nectria haematococca*) where positive fitness traits have been reported. In chives [233–235] and ryegrass [236–238], B chromosomes offer individuals of natural and experimental populations a selective advantage to survive under increasing stress conditions, while in *Nectria haematococca* Bs contribute to antibiotic resistance and pathogenicity [239]. Furthermore, a selective neutrality of B chromosomes has been reported by Porto et al. [220] in the whiptail catfish and Camacho et al. [208] in the grasshopper *Eyprepocnemis plorans*, whose fitness is not affected if Bs are present in low numbers.

At cell level, B chromosomes affect the physiology of the nucleus by generating numerical chromosome polymorphism. In rye, for example, mostly two to four B chromosomes can be found in natural populations, which adds further 580 Mbp for each single B to the basic genome size of 8.1 Gbp. With every additional B chromosome the cell size increases, leading to a longer duration of the mitotic cell cycle and a decline of the nuclear proteins and the RNA level relative to the number of Bs [189, 192]. In addition, delayed DNA replication of B chromosomes compared to the A complement replication were reported in *Brachycome dichromosomatica* [240, 241], the north shore marsupial frog (*Gastrotheca espeletia*) [242], the rat [243], the fox [244], and the fish *Astyanax scabripinnis* [245].

1.3.5. Molecular composition and organisation of B chromosomes

Beside phenotypic effects, inheritance mechanisms, and origin the molecular nature of B chromosomes fascinate, too. Until recently, technical limitations impeded the detailed analysis of B chromosome's molecular composition. DNA separation and extraction of the B chromosome proved to be difficult and few or no B-derived sequences were available. Hence, earlier studies resorted to techniques like density gradient centrifugation [246], renaturation kinetics [247], *in situ* hybridisation [248], and comparative digestion of genomic DNA with and without B chromosomes by using restriction enzymes [249]. Lately, technological advances, like microdissection [250] and chromosome flow-sorting [251], have allowed the isolation and sequencing of

1. Introduction

B-derived DNA providing a better insight into B's organisation and components. The knowledge gained about their components may also enhance the understanding of the Bs origin and evolution.

First screenings of the B chromosomes reveal that a major part of their sequence is similar to sequences of the A chromosomes, further underlining the As as a source of their origin. It has been equally shown that Bs have an affinity to accumulate high amounts of repetitive and mobile elements, such as tandem repeats, DNA-transposons and retrotransposons [211, 217, 221, 252, 253]. Since selection pressure is absent on Bs, they most likely lack functional loci [254] and that renders them to safe spots for repetitive elements. The tandemly repeated DNA sequences may diverge independently and are often associated with the formation of complex and highly heterochromatic regions on B chromosomes. The conglomerate of repetitive sequences localised on B chromosomes is usually higher in content to that of the standard chromosomes and shows discrepancies in its quantitative composition [204, 248, 249]. All identified repetitive DNAs, even the B-specific ones, share homology with different polymorphic A chromosome sites at different copy number levels.

In *B. dichromosomatica* [221] and *Drosophila subsilvestris* [255], the micro B chromosomes are entirely organised of tandem repeats, while in the grasshopper, rye, and maize both, satellite DNA and dispersed repeats, occur [217, 253, 256]. The tandemly repeated sequences are organised within blocks and located at the B subtelomeres and/or centromeres, as it has been shown in rye [249, 257], maize [256, 258], the wasp [259], the greater glider *Petauroides volans* [260], and the grasshopper [217]. As a reason of the strong repeat family enrichment on B chromosomes it has been suggested that the repeats may be involved in the maintenance and positive selection of Bs in the host species [261]. In the rat [243], the fox [244], and the fish *Astyanax scabripinnis* [245] the results indicate that the repeat composition of Bs delay their DNA replication compared to the standard karyotypes.

Additionally to the repetitive sequences conserved in both A and B chromosomes, B-specific sequences were found in several species. In rye two families of highly repeated sequences, *D1100* [249] and *E3900* [257], have been reported on the long arm of its B chromosome. Taking into account that they are transcriptionally active [262] and that they are located in a region relevant for the nondisjunction process [248], they are likely to contribute to the control of the nondisjunction mechanism. The B-specific sequences detected in maize, namely *pZmBs* [256], *pBPC51* [258], and *StarkB* [252], have been shown to be transcriptionally active, too. It has been assumed that they are functional components of the maize B centromere due to their centromeric location. The function of these transcribed B tandem repeats and their transcription mechanism remain unknown and need further investigations. Furthermore, tandemly repeated B-specific sequences have been described in *B. dichromosomatica* (*Bd49* [263] and *Bdm29* [221]), the wasp [264], the cyprinid fish *Alburnus alburnus* [203], the greater glider [260], and the raccoon dog *Nyctereutes procyonoides* [265].

Although repetitive DNA sequences represent a substantial component of the B chromosomes, the questions arise as to Bs contain genes or not and if yes, whether they are transcriptionally silent or active. So far, there has not been any evidence of B chromosomes carrying functional single copy genes yet. The lack of unique genes is striking, when one considers that some Bs

have an effect on their host phenotype (see section 1.3.4). It has been suggested that the detection of B-located genes, if available, is impeded by the massively amplified repetitive DNA and the high similarity of B sequences to A sequences. From previous studies contradictory results have been reported, suggesting the existence of active B-located genes based on their phenotypic effect on host species [266] or on the control of their own drive mechanism [267]. Putative B-located genes have been observed in mammals, like the yellow-necked mouse (*Apodemus flavicollis*), the red fox (*Vulpes vulpes*), and the raccoon dog (*Nyctereutes procyonoides*), and in the plant pathogenic fungus *Nectria haematococca*. The B chromosome of the pathogenic fungus comprises active genes that confer antibiotic resistance against antimicrobial compound produced by its host [239]. For the yellow-necked mice carrying B chromosomes an increased transcriptional activity of three genes, namely *CCT6B* (chaperonin containing *TCP-1*, subunit 6b (zeta)), *FHIT* (fragile histidine triad gene), and hypothetical *XP* gene, has been shown [268]. It remains unclear if the altered expression is a consequence of either additional active gene copies situated on Bs or of the presence of Bs in this species. The proto-oncogene *C-KIT* has been found in the B chromosomes of both the red fox and the raccoon dog, showing a high sequence and an exon-intron boundary conservation between them [269]. Here, too, the functional significance and the transcriptional activity of the B-located *C-KIT* gene has not been analysed yet.

The occurrence of multigene families (ribosomal RNA and histone genes) has been frequently described on B chromosomes. The ribosomal DNA sequences mostly arise in the form of nucleolar organising regions (NOR) and little is known about the role they play in B chromosomes. However, it is assumed that they may affect recombination and ensure the preferential transmission of Bs through the gametes [205]. Green et al. [254] have implied that the B-located NORs may alter the NOR expression of the A chromosomes, too. It has been shown that various B chromosomes, like that of *Plantago lagopus* (5S rDNA) [211], that of *B. dichromosomatica* (45S rDNA) [202, 270], or that of the fish *Haplochromis obliquoides* (18S rDNA) [271], are being enriched by a massive amplification of rDNA sequences. Due to the heterochromatic nature of Bs and their localization outside the nucleolus [272], almost all reported rRNA genes are inactive [211, 270–274]. Exceptions have been found on Bs of smooth hawkbeard (*Crepis capillaris*) [275, 276], of the frog *Leiopelma hochstetteri* [218], the grasshopper *Oedipoda fuscocincta* [277], and of the black fly *Simulium juxtacrenobium* [278]. The active NOR unit located on the supernumerary chromosome segments in the grasshopper *Oedipoda fuscocincta* increases the host capacity to synthesize proteins, which could confer an adaptive advantage to the species [277]. Despite the huge amount of ribosomal DNA sequences identified on B chromosomes, the presence of rDNA is not mandatory and Bs without rDNA exist, as described in rye [267].

Another multigene family observed on B chromosomes is the histone family, which is known to be composed of tandemly repeated gene clusters. Histone genes are usually involved in gene expression regulation, chromatin condensation and decondensation. The identified B-located histone clusters, especially H3 and H4, are highly conserved and show a high DNA sequence similarity to the genes located on the A complement [279]. On the migratory locust the B-located histone genes H3 and H4 show a higher DNA sequence variation compared to their corresponding A genes and are hence rather functionally inactive [279]. In the grasshopper *E. plorans* B-located methylated NORs are inactive, reinforcing the assumption that methylation may be used to silence potential B genes.

1. Introduction

B-located histone modifications have been reported in both, plant and animals, namely in *B. dichromosomatica* [240], rye [262], and several grasshoppers, like *Locusta migratoria* [279], *Eyprepocnemis plorans* [217], and *Rhammatocerus brasiliensis* [280]. The H3/H4 histone complex placed on the B chromosomes of the *B. dichromosomatica* show underacetylation (H4) [240] and a reduced level of euchromatic methylation marks (H3) [241]. There is no evidence of transcription activity, suggesting that methylation and deacetylation being mechanisms of gene silencing [261]. In rye two histone marks, a heterochromatic (*H3K9/27me*) and an euchromatic (*H3K4me*), located on the region involved in the nondisjunction process of the B chromosome, have been reported, though little is known about their organisation [262].

B chromosomes are often heterochromatic, albeit entirely euchromatic Bs have been detected in *Allium flavum*, in the fish *Characidium zebra* [281], in the snail *Helix pomatia* [282], and in *Scilla vvedenskyi* [241, 283]. The amount of heterochromatin of the heterochromatic B chromosomes varies substantially between several species, though their heterochromatic content is similar to that of the standard chromosomes [190]. In the grasshopper [284] and hare foot plantain [211] the B chromosomes are totally heterochromatic, whereas in the plague locust (*Chortoicetes terminifera*) the heterochromatic regions occupy 80% of the B, in *Puschkinia libanotica* 60%, in maize 50% [215], and 28% in rye [189]. The heterochromatic structure of B chromosomes influence their density during interphase and early prophase of mitosis and meiosis (show a very compact state), and abet a late replication in the S (synthesis) phase which is completed after the DNA replication in euchromatin cease [189].

1.3.6. Maintenance of B chromosomes

The transmission of B chromosomes to the next generation is a complex, but well-balanced mechanism that allows a stable B accumulation frequency over several years in a population. The observed equilibrium may result from the balance between the accumulation mechanism (increase frequency) and elimination caused by the harmful effects on host fitness (frequency decrease) [205, 285]. On the other side, it can be interpreted as a dynamic system that is under continuously shift due to environmental conditions and the ongoing contest between A and B chromosomes [204]. The most frequent transmission pattern is an accumulation process, so-called drive, that increases the presence of B chromosomes in gametes at a higher-than-expected rate. The accumulation process is irregular, non-Mendelian and differs amongst species. Three different drive modes were described in the literature: the pre-meiotic drive, the meiotic drive, and the post-meiotic drive [286]. The feature of increasing their frequency through drive mechanism underlines the parasitic nature of B chromosomes [285]. However, exceptions are known where no drive could be detected and the B accumulation is achieved at low frequencies through the benefit they exert on host fitness [189]. One such example are the Bs of chives (*Allium schoenoprasum*) that confer their host a germination advantage under stress conditions, while their number stays steady within the seedlings [235].

The mitotic and meiotic behaviour of B chromosomes transmission differ from each other. Most species show a normal mitotic transmission in somatic cells, so that Bs are equally distributed over all daughter cells. Mitotically unstable B chromosomes drift from these transmission mode and lead to an abnormal separation (mitotic nondisjunction) that end in daughter cells

with different numbers of Bs [193]. These explain the B characteristic to be present in a variable number and sometimes only in specific tissues within an individual. It is also possible that B chromosomes get lost from somatic tissues, but survive in germ lines [189]. Mitotic nondisjunction has been described in both plants and animals. In plants Bs are absent in the roots of *Aegilops speltoides* and *Xanthisma texanum*, while in *Crepis capillaris* the number of Bs varies in aerial organs [189]. In animals the somatic variation is widespread amongst the testes of male members of grasshoppers [286, 287].

The meiotic accumulation causes variation in the number of Bs passed to the progeny, usually exposing an increased number of Bs in the offspring than expected by Mendelian heredity [189, 193, 205]. The drive of B chromosomes is based on spindle asymmetry in some species. B chromosomes pass meiosis by lying outside the metaphase plate and are preferentially segregated to the spindle pole that contains the oocyte [191, 286]. For example, the mottled grasshopper *Myrmeleotettix maculatus* shows fluctuate transmission rates, depending on its gender: through females the transmission rate is higher (0.9) than through males (0.3) and both deviate from the Mendelian one (0.5) [286].

In flowering plants the most common drive is the post-meiotic drive based on directed nondisjunction in the gametophyte stage of the plant life cycle [191, 286]. Nevertheless, the life cycle stage at which the directed nondisjunction take place differ among grasses. The directed nondisjunction can take place at the first pollen-grain mitosis (e.g. fescue grass *Festuca pratensis*), at the second pollen mitosis (e.g. maize), or in two stages: at first pollen mitosis and at first egg cell mitosis (e.g. rye) [189, 191]. In maize, Roman et al. [288, 289] showed that the B chromosomes undergo nondisjunction at the second pollen mitosis resulting in two gametes, only one of them carries B chromosomes. The B-containing gamete shows preferential fertilisation, that may be explained by their position in the sperm nuclei that differ to the As ([289, 290], reviewed in [191]). While the B chromosomes regulate the nondisjunction process by themselves [291], González-Sánchez et al. [292] show that the selective fertilisation of the eggs is controlled by A-located genes. However, it remains unclear how B chromosomes manage to survive the intragenome conflict and how their position manipulates the critical fertilisation process to favour B-containing sperm [191].

Besides the nondisjunction process, B chromosomes maintain several other drive modes, too. In the grasshoppers *Eyprepocnemis plorans* and *Chorthippus jacobsi* a weak drive mechanism has been noticed that does not result in an increased frequency of Bs in the offspring. The observed non-random gamete fertilisation occurs due to the preference of oocyte and sperm to fuse with gametes bearing alleles other than their own [293] ones. For that reason Camacho et al. [208] proposed a near-neutral evolution model for the B chromosomes of grasshopper. According to this evolutionary theory the B chromosomes are endangered with extinction, since they influence the host fitness negatively and A-located genes suppress their accumulation [208]. Nonetheless, B chromosomes escape extinction by establishing new B variants, which outwit the A-located suppressor genes for awhile by surrogating the previous variant.

The drive mechanisms of the parasitic jewel wasp *Nasonia vitripennis* [294] and hermaphroditic flatworm *Polycelis nigra* [207] reveal the parasitic nature of the B chromosomes. The simultaneously hermaphroditic flatworm *P. nigra* can reproduce either sexually or asexually through

1. Introduction

gynogenesis carrying Bs in both the female and male line [207]. The asexually reproducing female line needs the sperm to trigger the development of the eggs, but the male genetic material is eliminated after cell fertilisation. In this regard, it is to be expected that B chromosomes present only in the male line would be expelled from the fertilised eggs, too, and the overall B frequency in the population would decrease. Remarkable is that B chromosomes are able to prevent the expulsion from the eggs and thus gain a selective advantage of the biparental transmission in an asexually reproducing host [207]. The paternal inheritance may explain the observed high B chromosomes frequency in the population of *P. nigra*. A similar case of paternal heredity has been described in Amazon molly *P. formosa* by Scharfl et al. [224], too. The Amazon molly is an all-female species that use the sperm of close related males to activate the egg fertilisation after which the paternal genetic material is turned out. Nevertheless, the offspring show sometimes microchromosomes that contain a parental gene responsible for black pigmentation. The incorporation of paternal DNA in an asexually reproducing organism seems to compensate for the disadvantages of asexuality and retains the species from extinction [224].

The jewel wasp *Nasonia vitripennis* underlines another particular case of an unusual transmission drive which is initiated by its supernumerary chromosome, the paternal sex ratio (PSR). The sex determination is acquired in wasps by the number of chromosomes which an individual bears: fertilised eggs lead to diploid females, while unfertilised eggs develop to haploid males [294]. The fertilisation of eggs by males carrying the PSR chromosome leads to elimination of the parental chromosomes excepting PSR. The diploid females become haploid males and increase the frequency of the supernumerary chromosomes in the population. The PSR frequency within the population is controlled by the reduced number of females and the competition among PSR carrying males [294].

1.3.7. The B chromosomes of rye

The B chromosome of rye (*Secale cereale*) is one of the best-studied B chromosome in plants and has been mentioned first in 1924 by Gotoh [195]. Rye (1C ~8,100 Mbp) contains between zero and eight mitotically stable B chromosomes, each having half the size of the normal A chromosomes (580 Mbp, see Fig. 1.6 B) [189, 295]. The rye plants can tolerate up to four B chromosomes showing no phenotypic effect, whereas Bs in a higher number decrease the plant vigour and fertility [296–298] remarkably. The most common form of B chromosomes in rye is the standard form, although morphological variations, like telocentric chromosomes and metacentric isochromosomes, exist (see Fig. 1.6 A). The different morphological types are likely to arise by deletion, centric misdivision and isochromosome formation and are usually rare due to their defective transmission or deleterious effects, like sterility, on the host [189].

Amongst the genus *Secale* B chromosomes have been found in both cultivated (*Secale cereale* ssp. *cereale*) and weedy (*Secale cereale* ssp. *segetale* and *Secale ancestrale*) forms of rye (*Secale cereale*) [300]. The B chromosomes among the mentioned rye taxa appear to be homologous suggesting a monophyletic origin [300]. This is remarkable since one would rather expect an increased rate of mutation and therefore different evolutionary fate. In breeding cultivars the B chromosomes are absent, since they are eliminated during selection due to their negative effect on plant fertility.

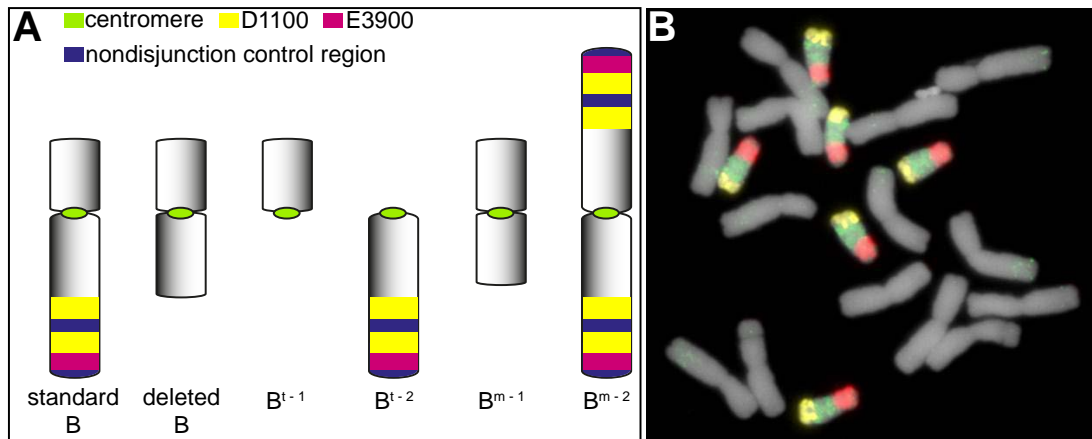


Figure 1.6.: **A)** Schemata of the chromosomal polymorphism in rye B chromosomes. The figure shows the standard rye B chromosome and five B chromosomes abbreviations found in individual rye plants. The five B structural variants include telocentric (B^{t-1} , B^{t-2}) and metacentric (B^{m-1} , B^{m-2}) chromosomes, as well as a chromosome that lacks the nondisjunction control region. In rye the nondisjunction control region (blue) is located at the terminal part of the long chromosome arm and contains the B-specific repeats *E3900* (magenta) and *D1100* (yellow) (figure adapted from Jones et al. [189]). **B)** Fluorescence *in situ* hybridisation (FISH) of B chromosomes on cultivated rye from Japan using three B-specific repeat probes: *ScC111* (red), *Sc36c82* (green), and *D1100* (yellow) (figure reprinted with the courtesy of Dr. Andreas Houben, [299]).

Rye B chromosomes are widespread in Asia, Europe and Northern America showing various frequencies between populations [301]. The highest frequency rates have been observed in Japan and Korea where it ranged from 19% to 92% [302]. In addition, the rye B chromosomes appear to have various effects amongst different genotypes, too: Bs show less severe negative effects on Korean rye populations than on other rye populations [302]. Rye B chromosomes influence the plant fertility depending on their transmission rate and/or on their number present in the maternal parent [192, 303]. A low B transmission rate implies a higher plant fitness, while high transmission decreases fertility: in 2B plants fertility is reduced by 25%, while in 4B plants it achieves 75% [192]. The fertility is considerably affected by the number of B chromosomes in the mother plant, too. A 4B mother plant can lead to fertile 0B progenies, whereas the 4B progenies of 0B mothers are sterile. These results suggest that Bs may control their own transmission and survival by promoting A alleles with a higher tolerance to their harmful effects [192].

The rye A and B chromosomes are of very similar DNA composition, having similar GC content [246] and repetitive DNA amount [247], except for one region located at the end of the long arm [216, 248]. This terminal region is heterochromatic, spans roughly 28% of the whole B chromosome and replicates late in the S phase [301]. It contains sequences shared with the

1. Introduction

standard chromosomes [216], but also strongly enriched B-specific sequences, namely *D1100* [249] and *E3900* [257]. The B-specific repeats reveal a complex organisation, *D1100* consisting of two repeat clusters separated by a small gap, while *E3900* is located toward the telomere [248, 253].

The B accumulation mechanism in rye (directed nondisjunction) is unique in plants and was first cytologically described by Hasegawa in 1934 [304]. B chromosomes undergo directed nondisjunction at post-meiotic mitosis during both male and female gametogenesis (see Fig. 1.7). The nondisjunction is triggered by sticking sites on either side of the B centromere that prevent normal separation of the chromatids at first pollen mitosis anaphase [305, 306]. The nondisjunction mechanism seems to be controlled by the terminal part of the long arm that might supply vital functions for this process [248, 305, 306]. Despite the lack of genes in this region a trans-acting element appears to be located there [190, 301]. The absence of this control region on the rye B chromosome inhibits its nondisjunction if no other intact B chromosome is present in the same nucleus. The autonomous nature of drive remains steady, even if the rye B chromosome is transferred into wheat [307] or into *Secale vavilovii* [308]. It seems reasonable to suppose that the B-specific repeat families, *D1100* and *E3900*, located in this region are involved in the nondisjunction process. This assumption is corroborated by the transcriptional activity that have been shown for both repeats in anthers [262].

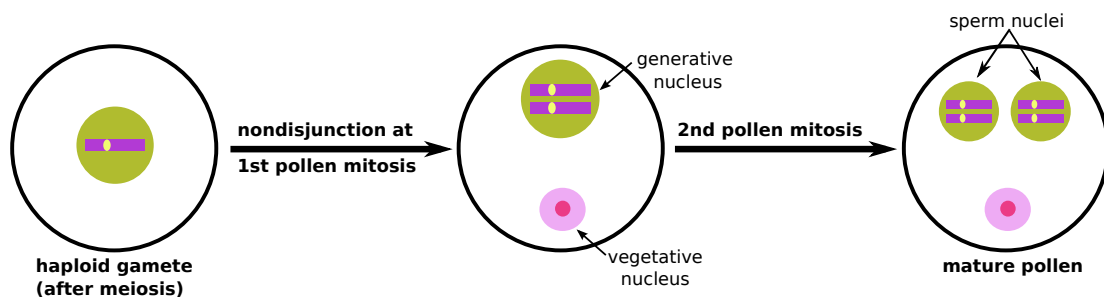


Figure 1.7.: The drive mechanism of the rye B chromosome in gametophytes. The rye B chromosomes undergo directed nondisjunction at the first pollen grain mitosis and migrate preferentially to the generative pole. As a consequence, the number of B chromosomes is duplicated by both sperm nuclei (figure adapted from Jones et al. [192]).

2. Material and methods

The *GenomeZipper* is a synteny driven approach. It has been developed to infer an approximate gene order for grass genomes that lack physical maps. The concept exploits syntenic conservation among grasses [309] and dense genetic marker maps of the species under investigation to position genes located in conserved regions along the chromosomes or chromosome arms. The gene order in the syntenic intervals along the extended marker backbone is transferred from the orientation in the respective reference genome. The *GenomeZipper* outcome is a linear ordered gene map which includes a marker scaffold, conserved synteny information from at least one model grass genome, and associated sequence data. Additionally, full-length cDNAs, expressed sequence tags (ESTs), and other data of interest can be placed in the linear gene map.

The *GenomeZipper* approach has its limitations and its outcome should be seen as a complementary resource and a surrogate for the complete genome sequences. The accuracy of the virtual ordered gene map is dependent on good and high-resolving genetic marker maps. The higher the marker density, the better the resolution of the linear gene map. The virtual ordered gene maps enable the development of a gene-based marker, the detection of structural rearrangements and complex relationships of synteny and collinearity among grass genomes, as well as comparative analyses of the conserved gene space in cereals.

Initially developed to arrange and orient the barley survey sequences [98, 99] along a genetic marker backbone, the *GenomeZipper* approach comes into use for several plant genomes (*Lolium perenne* [168], *Festuca pratensis* [170], *Triticum aestivum* [100, 310], *Aegilops tauschii* [311], and *Secale cereale* [102, 295]) and data types (454 sequences, contigs, or scaffolds) from individual chromosomes or chromosome arms.

The *GenomeZipper* pipeline is implemented as a semi-automatic process which includes three steps: repeat masking, detection of syntenic conserved regions and the integration ('zipping') step. The pipeline is managed by a wrapper checking the global configuration file, creating step specific configuration files, running homology searches against the reference genomes and controlling the flow of the individual program calls. Each step can be run independently, or as a combination of steps. Before running the integration step a manual validation of the automatically extracted syntenic conserved homologs is highly recommended. The verification is required to avoid the non-consideration of conserved blocks due to stringency parameters or to the selection of short conserved regions composed of protein kinase domains but not located in syntenic context.

To run the program a global configuration file in YAML (yet another multicolored layout) format [312] is required. The file (see Listing A.1) configures the initial settings for the repeat masking, conserved synteny detection step, and the 'zipping' step. The *GenomeZipper* output is saved in three different file formats: tab-delimited (.tab), comma-separated values (.csv) and

2. Material and methods

Excel (.xls). Basic statistics regarding the virtual ordered gene map are printed in an additional file. For an independent run of the individual steps local configuration files are required (see Listing A.2 and A.3). The individual *GenomeZipper* workflow steps are described in greater detail below (see Fig. 2.1).

2.1. Repeat masking

Grass genomes comprise a significant portion of repetitive DNA sequences that are scattered throughout the genome. Due to their high sequence similarity and dispersion they impede a complete genome assembly. To reduce the computational effort in gene space estimation the repetitive DNA content is identified based on sequence similarities between the NGS data and the MIPS-REdat Poaceae repeat library [314]. The alignments are performed using Vmatch [315]. The matches are selected according to their length (-l 100), identity (-identity 70), exdrop (-exdrop 5), seed length (-seedlength 14) and e-value (-evalue 0.001). All substrings covered by a match are masked with N characters and sequences containing more than 90% of masked bases are filtered out of the data set.

2.2. Identify conserved linkage blocks among grasses

The gene number present in rye [102] are estimated by sequence comparisons (BLASTX) of the low-copy 454 sequence reads against the protein sets of barley [316] and three reference genomes: *Brachypodium distachyon* [79], *Oryza sativa* [75], and *Sorghum bicolor* [80]. All homologs with at least 30 amino acids alignment length and 85% (barley), 75% (*Brachypodium*), or 70% (rice and sorghum) similarities are considered. The homologs preserved in conserved linkage blocks are identified by using a sliding window approach (windows size 0.5 Mbp, windows shift 0.1 Mbp). For each window the density of homoeologous matches (number of tagged genes divided by the sum of all genes) is calculated and regions, build up by multiple consecutive windows with a high degree of conserved synteny, are selected.

2.3. *GenomeZipper*

The *GenomeZipper* step consists of two main components: data preprocessing and data integration (“zipping”, see Fig. 2.2 A). First, all available data sets, such as species-specific genetic markers, NGS data, ESTs, full-length cDNAs, and syntenic conserved orthologous genes obtained in the previous step, are searched against each other for homology using BLAST (see Fig. 2.2 B). The reported matches are filtered for both, first-best hits and best bidirectional hits with a minimal alignment length of 30 amino acids or 100 base pairs and a minimal alignment identity. The alignment identity thresholds are applied in dependency of the evolutionary distances between the species to which the data sets belonged. The identity cut-offs used to filter homologous matches between various data sets in the construction of the rye and barley zippers are given below:

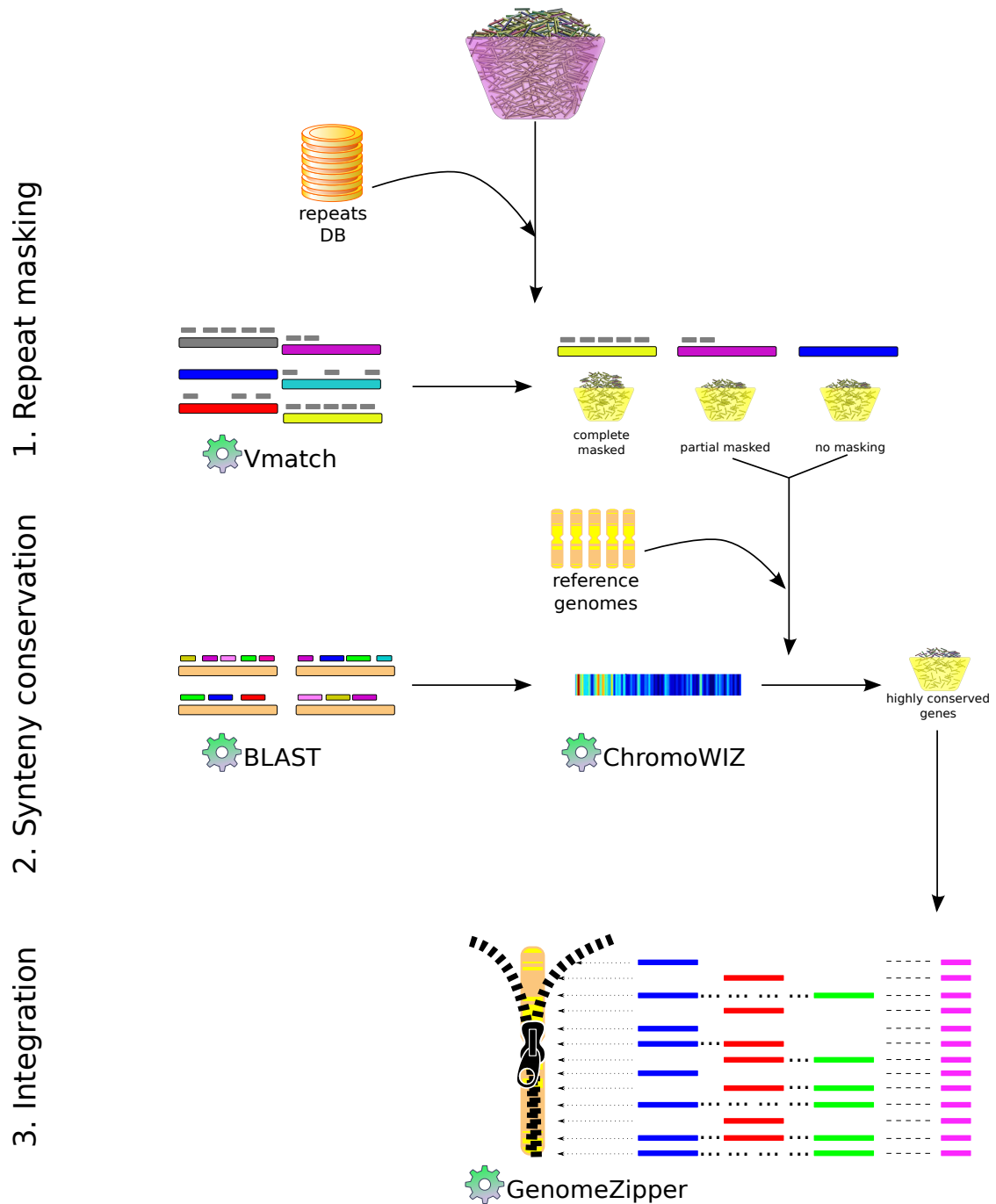


Figure 2.1.: Workflow describing the *GenomeZipper* pipeline, which consists of three individual steps that can be run independently: repeat masking, detection of syntenic conserved regions, and the integration of all data sets into a virtual linear gene map. See sections 2.1, 2.2, 2.3 for detailed description of the individual steps (figure adapted from Spannagl et al. [313]).

2. Material and methods

species	<i>Brachypodium distachyon</i>	rice	sorghum	barley	rye
barley	75%	70%	70%	95%	85%
rye	75%	70%	70%	85%	95%

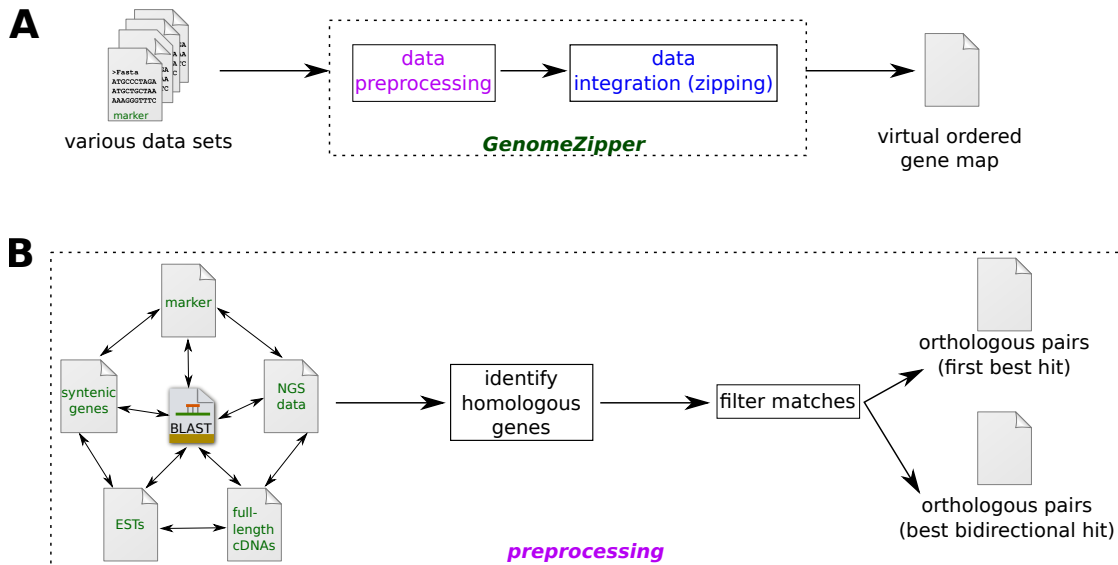


Figure 2.2.: Flowchart illustrating the *GenomeZipper* step: data preprocessing.

Second, the preprocessed data sets were interlaced in a virtual ordered gene map based on joint marker associations and best bidirectional hit classification. The genetic markers sorted in ascending order build the first layer of the new gene map. Progressively the syntenically conserved orthologs from three individual reference genomes are anchored to the marker scaffold. The integration starts with the collinear genes of the evolutionary closest reference genome (see Fig. 2.3 A) and proceeds as follows:

- (i) all collinear genes associated with a marker by best bidirectional hit are selected and placed alongside to the marker in the scaffold.
- (ii) if no unanchored collinear genes have remained in the previous step, then continue with the integration of collinear genes of the second evolutionary closest reference genome. Otherwise, determine the precise genomic position (Mbp) for both groups of collinear genes, with and without marker association, and construct a distance matrix between them.
- (iii) assign the remaining conserved genes to an anchored collinear gene according to the minimal distance between their genomic positions (closest neighbours).
- (iv) extend the scaffold at those places where the anchored collinear genes are affiliated to genes without marker association. Therefore divide the unanchored genes into clusters based on their genomic position and insert them into the backbone before and/or after the current anchored gene. The clusters insertion order is defined by the genetically ordered markers, while within the clusters the gene order is transferred from the order found in the respective reference genome. Thereby a sliding window approach is applied as follows:

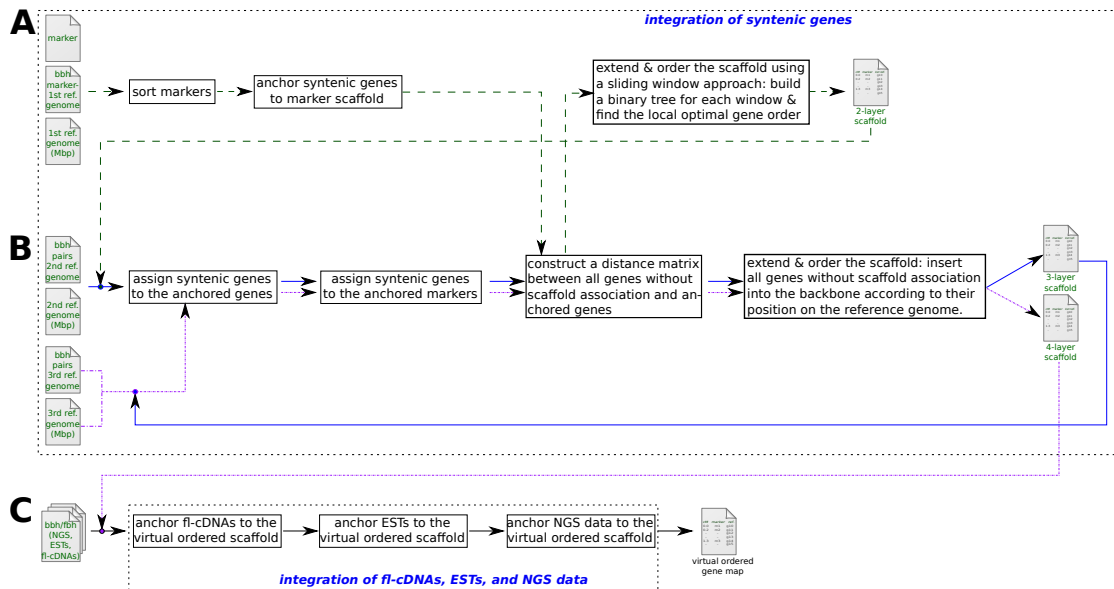


Figure 2.3.: Flowchart illustrating the *GenomeZipper* step: data integration (“zipping”).

- (a) define the window size by selecting the current anchored gene and n additional anchored genes located directly adjacent before and after the actual position in the scaffold (high n -values adversely affect the algorithm runtime, $n = 5$ has shown to be a good compromise between finding a good order quality and the algorithm runtime). If the current gene is the first or the last in the scaffold, then consider only the n genes located after or before this position, respectively.
- (b) determine the genomic positions (Mbp) of all anchored and unanchored genes in the defined window region (interval).
- (c) build a binary tree containing the genomic positions of all genes located within the selected interval.
- (d) investigate the gene order in the scaffold interval (ascending or descending). Therefore search through the tree using preorder traversal to find the path containing the optimal gene order for the examined interval (local optimal path). The local optimal path is the path where the average distance between the genomic coordinates of the gene candidates is minimal.
- (e) slide the window across the scaffold by one position and iterate through the (a)-(d) steps until all anchored genes are considered. The generated chromosomal backbone contains all unanchored syntenically conserved genes inserted before and/or after the anchored positions in ascending and/or descending order.

Next, the virtual ordered 2-layer scaffold is being extended by the collinear genes inferred from the second and third reference genomes (see Fig. 2.3 B). The procedure is the same for both data sets and will be described only once below:

2. *Material and methods*

- (i) go through the ordered scaffold and verify if there is a best bidirectional match between the already anchored genes and the new data set of collinear genes. If matches exist, then include the corresponding genes to the map.
- (ii) search the remaining conserved genes for best bidirectional matches with the marker and include the hits to the map.
- (iii) if all collinear genes are located to the virtual ordered gene map, then continue with the integration of additional information, such as full-length cDNAs, ESTs, and NGS data. Otherwise, construct a distance matrix between the remaining and anchored genes of the same species and calculate the minimal distance between each unanchored conserved gene and the genes included into the map (closest neighbours).
- (iv) sort the unanchored genes in ascending or descending order and insert them into the backbone before and/or after the position of the genes associated with them. The insertion order is deduced from the orientation of the two immediate neighbours (previous and next anchored gene in the map).

After inserting all collinear genes, each putative gene locus is supported by at least one representative, namely a marker and/or one or multiple syntenic genes. However, additional support evidences, such as full-length cDNAs and ESTs, can be added to the ordered gene map (see Fig. 2.3 C). At each gene locus, the anchored elements are surveyed regarding best bidirectional hits to full-length cDNAs and if they match they are inserted alongside the backbone. The ESTs are assigned to the scaffolds following the same procedure, except that the matches are selected using a first best hit criterion. In the last step the next-generation sequencing data is attributed to the chromosomal scaffold. Therefore, each element anchored at a gene locus is verified regarding matches (best bidirectional hit/first best hit) to the NGS data and anchored to the backbone. Redundant matches are entered only once. With the addition of the NGS data the integration of heterogeneous data sets and the construction of the putative ordered gene map is completed.

3. Publication summaries

All publications included in this thesis illustrate the use and application of the *GenomeZipper* tool on NGS technologies and comparative analyses based on the sequences of the barley and rye genomes, as well as the rye B chromosomes. The results give valuable insights in the genome structure and organisation of *Hordeum vulgare* and *Secale cereale*, and hints towards the speciation and evolution of the rye genome including its B chromosome.

The design, implementation and application of the *GenomeZipper* approach constitute the backbone of these articles. Applied on highly complex cereal genomes, like barley and rye, it is an effective strategy to analyse and structure the data obtained by NGS technologies. The construction of virtually ordered gene maps enable the exploration of genome structures and comparative analyses of the conserved gene space among grass genomes.

The summarised publications were published in peer-reviewed journals and can be found attached in appendix B. Further publications related to the topics of this thesis are listed on page v.

3.1. *Gene content and virtual gene order of barley chromosome 1H*

Mayer K.F., Taudien S., Martis M., Šimková H., Suchánková P., Gundlach H., Wicker T., Petzold A., Felder M., Steuernagl B., Scholz U., Graner A., Platzer M., Doležel J. and Stein N. 2009

The first article summarised in this thesis, *Gene content and virtual gene order of barley chromosome 1H*, was published in Plant Physiology in October 2009. It describes how a flow-sorted, low-pass shotgun sequenced barley chromosome (1H) and syntenic conservation among grasses can be used to construct a linearly ordered gene inventory. The high-resolution gene map is complementary to existing resources (genetic and expressed sequence tag maps) and a step towards developing a complete barley reference sequence.

Barley (*Hordeum vulgare*), a member of the *Triticeae* tribe, is the fourth important cereal grain in the world¹. Due to its large genome size (5.1 Gbp) and high repetitivity (over 80% repeats) it was previously economically unfeasible to generate a whole chromosome or genome sequence. The availability of NGS has changed this and NGS driven attempts to generate a WCS/WGS were started. The complete reference sequence is a premise to assess its gene space, to annotate the genes and to explore its gene diversity in cultivars and wild relatives. With the advent of next-generation sequencing technologies and the possibility of sorting chromosomes

¹<http://faostat.fao.org/>, September 4, 2014

3. Publication summaries

by flow cytometry, the analyses of such large genomes have become feasible. NGS provides the opportunity to generate massive data in a cost efficient manner and to gain knowledge about the genome structure of grasses on a much larger scale than it has been possible with earlier approaches. Despite the new achievements, assembling and ordering the huge number of short reads obtained by NGS to genomic scaffolds still remain a challenge.

To reduce the complexity of the barley genome, one individual chromosome (chromosome 1: 1H) was isolated by flow cytometry and shotgun sequenced at 1.3-fold coverage using the Roche 454 GSFLX platform. The chromosome 1H was selected due to its small size (~622 Mbp), which is equivalent to approximately one twelfth of the entire genome and because it can be easily direct sorted from the other six chromosomes. Additionally, sequence data derived from pooled, sorted barley chromosomes 1H to 7H (WCAall) was sequenced.

The repetitive content of both data sets was explored and similar proportion of repetitive elements found (77.5% in 1H versus 74.5% for the whole genome). The detected repeats were classified in two major groups: the retroelements (class I) and the DNA transposons (class II). The most frequent elements of the first class are the LTR retrotransposons, whereas the DNA transposon superfamily comprises the most frequent elements of the second class. Between the two, the LTR retrotransposons predominate, which is typical for plant genomes. Deviations in the repeat content of 1H and the whole genome were detected for CACTA elements (6% in 1H versus 6.4% for the whole genome) and ribosomal gene sequences (0.04% in 1H versus 0.13% in the entire genome).

To orient the 454 reads of barley chromosome 1H a new approach called *GenomeZipper* was developed. The method makes use of the good conservation of linear gene organisation (synteny) between barley, rice and sorghum, as well as a high-throughput genetic map to infer the order and position of the detected orthologous genes. The syntenic conserved orthologs between barley, rice and sorghum were identified. To make use of this the identified syntenic conserved rice (34%) and sorghum (43%) orthologs were positioned along the barley marker backbone to build a high-resolution, linearly ordered gene map. The number of genes localized on this particular chromosome and on the barley genome was estimated by using different complementary approaches. Averaging the results obtained by the different approaches, the number of genes on chromosome 1H is estimated to be 5,400 with a variation of $\pm 10\%$. Taking into account that the relative size of chromosome 1H is 12% of the complete genome, the total number of barley genes can be extrapolated to be about 45,000 with a variation of $\pm 10\%$.

The new hypothetical linearly ordered gene map gives insight into the structure and organisation of an entire *Triticeae* chromosome and positions its gene inventory along the chromosome. Since the approach is making use of fully sequenced reference genomes it also provides a virtual ordering of genes even in regions with low recombination frequency, such as centromeric and subcentromeric regions. However it cannot resolve local gene duplications and local rearrangements. The approach used and applied for the first time is a powerful approximation and complement existing resources.

My contribution to this paper was performing the bioinformatic analysis of the low-coverage sequenced barley chromosome 1H. Therefore, I developed and implemented the *GenomeZipper* method. The approach was applied on the barley data to assess and order its syntenic conserved

gene content. I estimated the entire gene amount by using the sequence similarities of the 454 sequence reads to the coding regions of two reference genomes (rice and sorghum), as well as to the EST resources from barley and wheat. In addition, I conducted the statistical evaluation of the raw and processed data and created 3 out of 5 figures. I also contributed to discussions and manuscript writing.

3.2. *Unlocking the barley genome by chromosomal and comparative genomics*

Mayer K.F., Martis M., Hedley P.E., Šimková H., Liu H., Morris J.A., Steuernagl B., Taudien S., Roessner S., Gundlach H., Kubaláková M., Suchánková P., Murat F., Felder M., Nussbaumer T., Graner A., Salse J., Endo T., Sakai H., Tanaka T., Itoh T., Sato K., Platzer M., Matsumoto T., Scholz U., Doležel J., Waugh R. and Stein N. 2011

The second publication, *Unlocking the barley genome by chromosomal and comparative genomics*, reports about the first genome-wide high resolution sequence-based gene map of the barley genome and in-depth comparative analyses with other grass genomes. The article was published in *The Plant Cell* in April 2011.

Barley (5.1 Gbp) and wheat (17 Gbp) are two of the most important crops in the world. Both lack a complete genome sequence yet, which has been impeded by the size and complexity of their genomes. Barley and wheat share extensive syntenic conservation. Since wheat has an additional layer of complexity due to its hexaploidy and thus the 3 fold genome size, barley can also serve as genomic model or blueprint for bread wheat. The barley chromosomes were sorted by flow cytometry and sequenced using Roche 454 technology. To order and structure the obtained sequences a bioinformatic approach, *GenomeZipper*, was applied. Compared to previous applications of the tool, the *GenomeZipper* approach was extended to integrate a third reference genome, as well as full length cDNAs and DNA hybridisation microarray data.

Using different barley specific data sets, like Roche 454 sequence reads, genes assigned by array hybridisation and full-length cDNAs, and three reference genomes (*Brachypodium*, rice and sorghum) a cumulative set of 24,698 non-redundant homologous genes was identified. Based on evaluations of discovery rate and sequence coverage an overall content of 32,000 genes for the entire barley genome was estimated. Roughly 68% of the estimated genes are localised in syntenic conserved regions and were arranged along the barley individual chromosomes and chromosome arms. The order is inferred from the barley genetic marker backbone and the gene ordering in the three grass genomes used as genomic models.

Due to the low recombination rate in centromeric regions the centromere positions of the seven barley chromosomes are limited in genetic resolution. Using the linear gene maps, we were able to identify the corresponding genomic position of the centromeres for all seven chromosomes and thus improved knowledge on the composition of centromeres. According to the gene maps 14% of the syntenically conserved genes are allocated to these regions.

The seven ordered gene indices revealed a higher syntenic conservation between barley and *Brachypodium* than between barley and rice or between barley and sorghum, reflecting the closer

3. Publication summaries

phylogenetic relationship between the two species. However, to overcome limitations imposed by species-specific differences it is important to use all three organisms as reference to order and structure the barley genome. The order and orientation comparison between barley, *Brachypodium*, rice and sorghum reveal numerous local rearrangements and nine duplicated genome segments. Six of the observed duplications belong to previously described grass specific segmental duplication, while three are barley-specific.

The comparative analysis between the ordered barley full-length cDNAs and wheat EST bin maps support the previous observed high syntenic conservation. Only three wheat chromosomes (4A, 5A and 7B) are involved in chromosomal translocations. The ordered full-length cDNAs also indicate that the barley genes evolve under strong purifying selection, whereas only few genes (105 fl-cDNAs) have been detected to be under positive selection.

The article concludes that the set of seven genome zippers are a step toward a complete reference genome sequence of barley. The genome zipper maps are proposed as reference models and surrogate for both the barley genome itself and for closely related *Triticeae* crops until further resources, such as physical maps, became available. Furthermore the *GenomeZipper* approach was suggested as helpful for the ordering of genes along wheat (*Triticum aestivum*) and rye (*Secale cereale*) chromosomes, as well as for other crop and legume genomes where individual chromosomes can be sorted by flow cytometry.

My contribution to this publication was performing most of the sequence analysis. I extended the *GenomeZipper*'s implementation to improve the accuracy of the gene ordering by including an extra reference genome. The advanced method also supports the integration of additional data sets, like full-length cDNAs and genes assigned by array hybridisation. The improved *GenomeZipper* approach was applied on all seven individual barley chromosomes. Furthermore, I statistically evaluated the raw and processed data (marker, 454 sequence reads, fl-cDNAs), identified and masked the repetitive 454 sequence reads, and I associated the barley fl-cDNAs to individual barley chromosomes. I also estimated the gene number in barley, analysed the syntenic conservation of this genome to wheat, created figures and contributed to manuscript writing.

3.3. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences

Martis et al. 2012

The article *Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences*, published in Proceedings of the National Academy of Sciences in July 2012, shows the origin and molecular make-up of supernumerary B chromosomes in rye.

The aim of our work was to elucidate the rye B chromosome enigma by using flow cytometry, next-generation sequencing and comparative analysis. Therefore the flow cytometric sorted A and B chromosomes of rye were shotgun sequenced by Roche 454, repetitive sequences were filtered out and comparative analyses of A and B chromosomes and several reference genomes

(*Brachypodium distachyon*, *Oryza sativa*, *Sorghum bicolor* and *Hordeum vulgare*) were performed.

Most grass species are known to contain large repeat-rich genomes. The rye genome (A chromosomes) and its B chromosomes are no exception. Over 90% of their sequence is composed of repetitive elements, but the A and B chromosomes show significant differences in repeat composition and frequency. The B chromosomes carry specific, long satellite repeats (0.9 - 4 kb) and sequences corresponding to *Bianka* family of *Ty1/copia* elements. In the rye genome over 70% of the repeats are represented by fewer than 60 different repeat families. Additionally, B chromosomes accumulate large amounts of mitochondrial and chloroplast DNA, which exceed the organellar DNA insertions in the A chromosomes.

To complete the picture of the B chromosome DNA composition its gene content was explored and surprisingly a high amount of gene fragments was found. A non-redundant gene count of at least 4,946 genic sequences was extracted and thus the general belief that B chromosomes do not carry genes was disproved. Regarding the completeness and functionality of the B-located genes no conclusions can be made with the available data. Due to the low sequence coverage of the B-genomic data (0.9-fold) an assembly of the gene-rich 454 sequence reads is not feasible and they can not be definitely categorised as transcriptionally active genes or pseudogenes. To make a statement about their transcription activity the generation and analyses of high-throughput transcriptome data and wet-lab validation (qPCR) are required.

Using the close syntenic relationships among the rye and barley genome we were able to trace the putative origin of the rye supernumerary chromosome. The B chromosome reads show a high similarity to segments of barley A chromosomes, especially to regions located on the chromosomes 2H, 3H, 4H and 5H. The tagged barley regions correspond to two chromosomal regions in the rye genome, namely to chromosome 3R and 7R. This allows us to propose a model of B chromosome evolution, including its origin by recombination of two A chromosomes (3R and 7R) followed by accumulation of additional A-derived sequences, organellar insertions and amplification of B-specific repeats. The origin of the selfish chromosomes is estimated to have occurred 1.1 - 1.3 Mya.

For this article I was in charge of performing synteny assignments of the rye B chromosome to barley and three model genomes (rice, sorghum and *Brachypodium*) and of unravelling the DNA composition of the B chromosome. The detection of B-genic sequences and organellar DNA altered our view of rye B chromosomes DNA composition, which were thought to carry no genes. I also elucidated its origin by tracing it back to two rye A chromosomes (3R and 7R) and proposed a model of rye B chromosome evolution. Furthermore, I contributed to figure creation, discussions, and manuscript writing.

3.4. Reticulate evolution of the rye (*Secale cereale* L.) genome

Martis et al. 2013

The last publication presented here, *Reticulate evolution of the rye (Secale cereale L.) genome*, reports the construction of linearly ordered gene maps for the rye genome and provides insights

3. Publication summaries

in its heterogeneous composition which indicates reticulated evolution as a result of introgressive hybridisation and/or allopolyploidisation events. The article was published in *Plant Cell* in October 2013.

Rye (*Secale cereale*) is one of the main cereals for food and feed in Eastern and Northern Europe. Its close relationship to wheat and barley as well as its stress tolerance to frost, drought and marginal soil fertility propose rye as a model for functional analyses and crop improvements. As a result of its genome complexity, huge genome size (8.1 Gbp), and only regional importance the analysis of the rye genome remain a challenge and lag behind other cereals.

A new bioinformatic strategy of genome analysis (*GenomeZipper*), involving chromosome flow sorting, in part low coverage next-generation sequencing, extensive syntenic conservation to model genomes, and a dense genetic transcript map, enabled to unravel the rye genome. Along a high-density genetic marker backbone a cumulative set of 22,426 rye genes was ordered and positioned. The virtually ordered gene maps provide fundamental insights into the organisation and structure of the seven individual rye chromosomes.

Compared to barley and wheat rye has undergone a series of rearrangements that led to a 'mosaic' genome structure. The six major translocation events observed involve six out of seven rye chromosomes. Most of them occurred after the split of wheat and rye lineages. Only 1R exhibits collinearity over the entire chromosome length. All other chromosomes are composed from 2 to 4 conserved syntenic segments of barley chromosomes.

In total 17 regions were found to differ between rye and barley. For each region the sequence similarity and conservation patterns were ascertained and the results revealed that they are not homogeneous through all segments. The unbalanced degree of syntenic conservation, the varying sequence homology, and phylogenetic observations in these segments demonstrate that the modern rye genome is a mosaic of segments of several ancestral genomes.

Based on these observations a revised model of genome evolution was proposed for the rye genome evolution and speciation. The detected interspecies rearrangements are evidence of reticulate speciation triggered by introgressive hybridisation and/or whole genome or chromosome duplications and may be a consequence of the outbreeding nature of rye.

Chromosome flow sorting and next-generation sequencing technologies enable unprecedented access to *Triticeae* genomic resources. This will allow detailed comparative analyses among cereals and their wild relatives and will potentially influence crop breeding and enhance our understanding of the dynamics of grass genome evolution and speciation.

For this publication I performed the majority of the data analyses: I applied the *GenomeZipper* on the rye chromosomes, assessed syntenic conserved regions between rye and four reference genomes (barley, rice, sorghum and *Brachypodium*), and I conducted comparative and statistical analysis between them. In addition, I determined the complex chromosomal rearrangements that occurred in rye and studied their role and implications in the rye genome evolution. Besides, I generated all figures, discussed and interpreted the results with co-authors and contributed to the manuscript writing.

4. Discussion and conclusions

The annotation and exploration of plant genome features and diversity has lagged behind that of other organisms, such as bacteria, fungi, and animals, due to their often striking complexity (see section 1.1.1) and, to large extends, the lack of fully sequenced genomes. Advances in flow cytometry to sort chromosomes and in sequencing technologies (see section 1.1.2) have recently enabled to tackle these genomes and overcome the complexity imposed by ploidy level and genome size. Although genomic sequences have become available, new challenges and difficulties have arisen in processing, assembling, and annotating these high throughput DNA sequences, due to the massive amount of NGS data, the high repetitiveness of plant genomes and computational limitations [77]. Therefore, there was a strong requirement to develop novel tools and methods to address the structure and organisation of large plant genomes. The research work carried out in this thesis can be divided into two major tasks. First, the design and development of a novel approach (*GenomeZipper*) (see Chapter 2), that aims to address some of the above mentioned constraints. The *GenomeZipper* provides genome scale ordering of the low-complexity regions of the low-pass sequenced barley, rye, and wheat genomes. Second, the use of *GenomeZipper*-generated sequence-based gene maps allowed to undertake comparative analyses with the aim to elucidate and understand their genome organisation and their evolutionary history.

4.1. Cereal genomes unlocked using the *GenomeZipper* method

Cereals contain impressively large and complex genomes, whose exploration enables the deciphering of their structure, organisation, and evolution. This understanding can help to increase the cereal production and help to develop new varieties with specific characteristics, like resistance to diseases and environmental stress. Up to now this has been impeded for the *Triticeae* tribe by the lack of genomic sequence information. This thesis provides an overview of the first low-coverage 454 shotgun sequencing (1-fold to 1.5-fold on average) of the flow-sorted barley and rye chromosomes/chromosome arms and a characterisation of their gene content. Both barley (5.4 Gbp) and rye (8.1 Gbp) genomes are 1.7 to 2.6 times larger than the human genome and contain more than 80% repetitive elements. Their pronounced genome complexity and largely incomplete genomes prevented scientists from their accurate assembly and required alternative approaches to access their genomic information. Based on the exploitation of the well-established synteny conservation among cereal genomes [154], it was possible to create linearly ordered, dense, sequence-based gene maps for each of the seven rye and barley chromosomes using the *GenomeZipper* method. The proposed set of *GenomeZipper* ordered genes

4. Discussion and conclusions

covers about two thirds of the estimated barley and rye genes and can be viewed as the closest ordered approximation to their reference genomes. The new genomic resources provide a high quality surrogate for working genomes and for closely related grasses until their physical maps and complete genome sequences become available [317].

The *GenomeZipper* concept has first been applied on barley chromosome 1H (see section 3.1). The *in silico* method provided a highly structured and ordered gene-map of chromosome 1H, positioning 1,987 genes along the chromosome. The gene inventory is based on the linkage of low-pass 454 sequence reads, synteny information derived from model genomes, and fl-cDNA [185] and EST (<http://www.harvest-web.org>) collections embedded in a framework of species-specific genetic marker maps [318]. The syntenic information used to project the putative order of the orthologous genes has initially been inferred from two reference genomes (rice and sorghum) and has been extended by a third model genome (*Brachypodium*). The addition of a third organism, which is evolutionary more closely related to the *Triticeae* allowed to refine and specify the ordering of the survey sequences along the chromosomes. The identified syntenically conserved regions among barley, rye, *Brachypodium*, rice, and sorghum confirm previous observations based on low resolution RFLP marker and transcript map data comparisons [156, 319]. In follow-up research efforts, virtually ordered gene inventories of 21,766 and 22,426 genes have been predicted for the entire barley and rye genomes (see the sections 3.2 and 3.4), respectively. The high-resolution gene maps allow to establish the linear gene order even in poorly recombining regions, such as the pericentromeres and centromeres, where previous genetic efforts to order genes have failed. Hence, for six out of seven barley genetic centromeres a precise localisation could be undertaken and 3125 genes were positioned therein. All but nine of these genes have been allocated to the proximal or distal chromosome arms and a linear order has been proposed for them. Although the *GenomeZipper* provides a rich source of information in these regions, the order accuracy is poorer than in other parts of the chromosome and may contain wrong ordered gene alignments. This potential shortcoming asks for novel genetic strategies, such as deletion mapping or genome-wide studies in diverse populations [320], to increase the recombination in the pericentromeric and centromeric regions.

The unlocking of the barley and rye genomes pave the way for not only accessing syntenic conserved regions, but also for estimating the whole gene content. The gene estimations have been based on stringent sequence comparisons of fl-cDNAs and 454 sequences against the sequenced grass model genomes. In barley, 77% of the used fl-cDNAs and 24,700 genes derived from 454 sequences and array-based data have shown a homologous match to the protein-coding genes of *Brachypodium*, rice, and sorghum. Considering this observations, a total set of ~32,000 genes has been estimated for the entire barley genome. In rye, based on the measured sensitivity (78.7%) and the ~31,000 genes with matches to the homologous reference genes mentioned above, ~40,000 genes have been postulated for the whole genome. The revealed gene counts are in the same magnitude as reported for other grass genomes, such as rice (~39,000) [75], sorghum (genome annotation v1.4: 27,640) [80], and *Brachypodium* (genome annotation v1.2: 26,552) [79]. However, to this end the limited sequence coverage and the abundant number of gene fragments and pseudogenes that have been found for *Triticeae* genomes [64, 316] are considerable shortcomings and limitations.

What started as a *in silico*, genome driven, experimental attempt to assess and order the gene

4.1. Cereal genomes unlocked using the *GenomeZipper* method

content of low-coverage sequenced cereal genomes has also proved to be a valuable tool for laboratory-based research and plant breeding. The *in silico GenomeZipper* approach is a very effective and powerful strategy to generate sequence-based gene maps not only for flow-sorted barley and rye chromosomes (see the sections 3.1, 3.2, and 3.4), but also for other cereal genomes. It has been successfully applied to several individual chromosomes and genomes, such as wheat chromosomes 4A [100], 4D ([321], accepted and in press), and 1BL [310], wheat homoeologous group 1 [64], wheat chromosome arm 4AL with *Triticum militinae* introgression ([322], unpublished), meadow fescue chromosome 4F [170], *Lolium* [168], *Ae. tauschii* [311], *Ae. umbellulata* ([323], unpublished), and wheat [324]. The accuracy of the predicted gene indices, and therefore also the *GenomeZipper* approach, has been validated for barley [325] and the long arm of chromosome 1BL in wheat [310]. The validation was performed by both, *in silico* comparisons and experimental testing. The *in silico* screening of wheat 1BL against its physical map revealed that 82.5% of the genes share the exact same order [310]. The inconsistencies (particularly inversions) observed for the remaining genes (17.5%) could be partially attributed to errors in the genetic map scaffold and partially to the low resolution of this map at certain positions (centromere), that make an orientation unreliable. For barley, Poursarebani et al. [325] evaluated the accuracy of the *GenomeZipper* strategy to more than 94%. To do so, the authors conducted genome-wide comparisons of the barley zipper maps against transcript-derived markers and a wet-lab experimental validation (fine-scale) of a small segment located on chromosome 2HL. The error rate (~5%) determined in this study is in accordance with that noticed in other barley consensus genetic maps [325]. The remarkable accuracy of the approximated order of the *GenomeZipper* gene maps makes them a valuable genomic resource to exploit genetic diversity of *Triticeae* crops and to promote their improvement. Hence, the ordered gene maps of barley and wheat have so far been used for:

- (i) marker development and positional cloning to facilitate marker-assisted breeding by selecting genes for particular traits (e.g. higher yield, disease resistance) [326–331],
- (ii) genotyping by sequencing,
- (iii) systematic anchoring of clones/contigs to physical maps and their validation [180, 332],
- (iv) identification and elucidation of chromosomal structures and collinear regions between grass genomes [100, 295],
- (v) comparative analyses of gene content and organisation [64, 333], and
- (vi) tracking back the origin of the rye B chromosomes to the standard rye karyotype and to the organellar genomes [295].

Although the *GenomeZipper* proved to be a very successful tool for gene isolation and genome-wide analyses, the method has limitations. The approach uses both genetic marker and conserved synteny information to order and position the survey sequences. Hence, its precision depends on resolution and accuracy of underlying genetic maps. Incorrect genetic maps lead to erroneous anchoring, as shown e.g. for wheat chromosome arm 1BL [310]. Small-scaled rearrangements, such as insertions, deletions, inversions, duplications, or translocations, result in discontinuities in collinearity and hamper a correct ordering, too. Interrupted synteny blocks have been observed among several grass genomes, like wheat, rice, maize, and sorghum [175, 334, 335].

4. Discussion and conclusions

Moreover, the assigned collinear blocks comprise only roughly 60% - 70% of the estimated gene set in cereal genomes [64, 99, 336], omitting all non-syntenic and species-specific genes. To address and order not only the whole gene space, but all available genomic sequences along individual chromosomes, other methods need to be applied, e.g., physical and genetic maps, or POPSEQ [337]. An additional constraint is the availability of genetic maps and the availability of close relatives with completely sequenced genomes required as a scaffold for establishing the virtual ordered gene maps.

The presence of contamination in next generation sequencing data is common and its removal represents a significant challenge. The contamination arises from bacteria, human, and organellar DNA, as well as from other chromosomes that were adjacent in the flow histogram [338]. Sequences contaminated with bacterial or human DNA present no problem for the *GenomeZipper* approach, since these do not match the filtering criteria. However, contamination with other chromosomes may impede the correct identification of syntenically conserved regions for the species of interest. This may affect an accurate alignment of genes along the chromosomal backbone. The *GenomeZipper* is largely insensitive to this type of sequence contamination, provided that the assignment of the collinear regions has been manually verified or that the contaminated reads/contigs are located on different conserved blocks.

Depending on the sequence resources used for integration into the virtual ordered gene maps, their use for resequencing projects may be limited. Low-coverage survey sequences (1-2x) cannot be used to generate a reasonable assembly and the individual 454 reads are too short to serve as a suitable reference for mapping short Illumina reads used for resequencing. Besides, assemblies derived from anchored ESTs or fl-cDNAs do not consider intergenic sequences and have only limited use for mapping genomic NGS reads. However, these approximated chromosome scaffolds can be used to establish evolutionary relationships between different species and to detect rearrangements or identify sequence contaminations. Contig or scaffold sequences derived from sequence assemblies cover a high proportion of the low-copy regions and thus are more useful for the above mentioned tasks, albeit restricted by the number of syntenic genes [339].

Due to its simplicity - but still high effectivity and utility - of the *GenomeZipper* approach, several research groups have adopted the idea behind the tool and implemented their own version to meet their needs. The created *GenomeZipper* clones were mostly applied on wheat, but were also considered for oat and garden pea. For wheat, the zipper similar strategies were applied on several individual chromosomes sequenced at low coverage (1.5x to 10x) with 454 and Illumina sequencing technology, namely on the chromosomes 5A [340], 1AL [332], 3A [339], and 3B [341], and on all chromosomes of homoeologous group 7 (7A, 7B, 7D) [342]. Furthermore, Alnemer et al. [343] presented a web application, the so-called 'Wheat Zapper', to calculate the collinearity between wheat, rice, sorghum, and *Brachypodium distachyon*. Beside the gene order prediction based on synteny, this tool enables primer design, prediction of intron/exon boundaries, and a tabular and graphic display of the results [343]. The authors claim a 65% accordance between their tool and the *GenomeZipper* approach, based on the wheat chromosome 5A gene map comparison. Since the predicted gene map of wheat 5A [340] has not been built using the *GenomeZipper* approach (but a similar strategy), the consistency of the 'Wheat Zapper' and *GenomeZipper* ordered gene maps has not been tested yet. Both tools differ from each other in their method to assess collinear regions, as well as in anchoring and ordering the genes

4.2. Insights into the structure, organisation, and evolution of cereal genomes

along a chromosomal scaffold. The *GenomeZipper* detects syntenic conserved regions using a sliding window approach that tracks dense blocks of homologous matches between query and reference genomes [313], while the ‘Wheat Zapper’ uses a majority consensus approach [343]. The precision of the gene order prediction in the ‘Wheat Zapper’ only depends on the syntenic relationships between the input sequences and the reference genomes. The order of the gene maps generated by the *GenomeZipper* is determined by a species-specific genetic marker backbone, followed by a step in which the order is deduced from the syntenic information. The ‘Wheat Zapper’ may be useful for the fast assignment of orthologous relationships among wheat sequences and the reference genomes, but it’s missing the ability to study large genomic sets of data and to achieve the exactness which allows the use of species-specific genetic maps. Further, the *GenomeZipper* can readily be applied to several plant genomes with full sequenced reference genomes available, while the ‘Wheat Zapper’ is restricted on wheat, rice, *Brachypodium*, and sorghum. Nevertheless, this study agrees with the *GenomeZipper* observations that syntenic conserved regions deduced from multiple reference genomes provide a high potential for gene identification and orientation in species that lack a complete genome sequence.

In addition, as part of this thesis, the previously described barley [98, 99] and rye [102] virtually ordered gene maps have been made accessible to the research community via the PGSB PlantsDB database [314]. The web presentation of the gene maps facilitates the search and navigation through individual positions along the chromosomes and links them to other genomic resources, such as SNP marker, cDNAs, ESTs, survey sequences, and reference genome sequences (see Fig. 4.1). In addition, the gene maps can be downloaded as bulk data files and are complemented by other data resources (e.g. physical maps), upon availability. Furthermore, the PGSB PlantsDB has been extended recently to provide access to the ordered gene maps of ryegrass [168] and wheat [324].

4.2. Insights into the structure, organisation, and evolution of cereal genomes

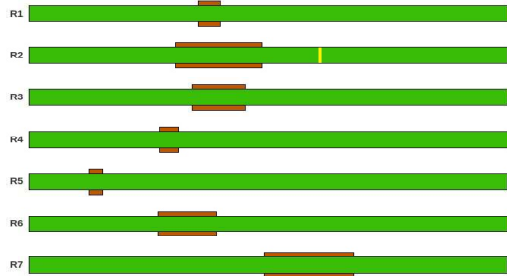
Grasses have been subjected to comparative genomic studies for over three decades, unveiling a remarkable conservation within orthologous chromosome segments. The first attempts to gain new insights into genome characteristics, structure, and evolution of related grass genomes were based on low resolution genetic markers (macro-collinearity), followed later by sequence-based comparisons (micro-collinearity, see section 1.1.4). The high genome collinearity established at global level using low resolution marker maps lead to the presumption that grasses can be considered as a ‘unified grass genome’ and displayed as a ‘circle model’ [154, 156, 309, 344]. Despite the good conservation, numerous small chromosomal rearrangements, like inversions, translocations, deletions, insertions, or duplications, have been detected at micro-synteny level [177, 334]. The frequent disruption of syntenic conservation cast partial doubt on the proposed ‘single syntenic genome’ model, suggesting that structural relationships between grasses are more complex than initially assumed. Although these genomic variations may limit the transfer of information about important trait genes from one species to another, the conserved synteny between grasses still provides reliable information (i) to identify and delimit syntenic linkage

4. Discussion and conclusions

GenomeZipper Table Overview

Choose a chromosome:

Choose directly your loci of interest by click on the designated region in the graphical chromosome representation (brown boxes highlight loci in centromeric regions):



Search

Search GenomeZipper for marker or syntenic genes by name

Your input: e.g. c13431 170-183-259, Xscm30, Brad14q00206.1, Sb06q001580.1

Search GenomeZipper for marker, syntenic genes, f1cDNAs, ESTs, Zipper reads or f1cDNA reads by name

Your input: e.g. G76ZXPNO2GFL35_7R, AK374103, Sce_Assembly03_c4565

GenomeZipper Table for chromosome R1

To change the loci of interest click on the desired region in the graphical chromosome representation (brown boxes highlight loci in centromeric regions):



Loc1 326-350 of 2506

Loc1	chr. Position	Marker	in syntenic relationship with			Link to		
326	-	-	Brad12g36730.1	Os05g0140100	Sb09g003100.1	f1cDNAs	Reads	ESTs
327	-	-	Brad12g36710.1	-	-	f1cDNAs	Reads	ESTs
328	-	-	-	Os05g0140600	-	-	Reads	-
329	-	-	-	Os05g0140600	-	-	Reads	-
330	-	-	Brad12g36687.1	Os05g0140600	Sb09g003150.1	f1cDNAs	Reads	-
331	-	-	Brad12g36670.1	Os05g0141300	Sb09g003170.1	f1cDNAs	Reads	-
332	-	-	Brad12g36660.1	Os05g0141400	Sb09g003210.1	-	Reads	-
333	74.26	c24346_153	Brad12g36647.1	Os05g0141500	-	f1cDNAs	Reads	ESTs
334	-	-	Brad12g36640.1	Os05g0141700	Sb09g003230.1	-	Reads	-
335	-	-	Brad12g36620.1	-	-	-	Reads	-
336	74.48	c9847_450-324	-	-	-	f1cDNAs	Reads	-
337	76.53	c22321_112	-	-	-	f1cDNAs	Reads	ESTs
338	77.27	TC76893-1R	-	-	-	-	-	-
339	77.27	c2960_331	-	-	-	-	Reads	ESTs

Sequence report

Name: GP3EJ2G02J16H6_1R

Type: 454read

Extraction from: Secale cereale

Length (bp): 483

Sequence for GP3EJ2G02J16H6_1R

Download:

```

CTGATTTCCATCTTGTACTCAAGCTCTTGAAGTCAITTTTACTCGAAACCTTTTCTTGGCTTAGAT
CTPATCTAGTGAATACCTGCAATGCTTAAGCTAAGCATATCCAGCCAGTTTGTATATATCTCACATGGCTCTG
TTAATAGAGAACTGTAGCTTGCOCGAAATGAGGCACTGCTTTTCATATCTCCCATGTTTATCCOCOC
TTTGTATAGCTGTGTGAGCCTGAGCACTGAGCACTGAGCACTGAGCACTGAGCACTGAGCACTGAGCACT
AACTCTTATAGCTTATGATGAACTGTCTATGTTGCAAAATTTATATACCCCTGATCCGTAATATGTGAA
ACCTTAATTAATACTGGACACTTATTATGATCGGAGGAGTACTGATTAATGCTAACAAAATTGACATAA
GAGAGATATTTCAATGATCTTATCTCTTCAGAGAAATGATATAGCAG
    
```

Figure 4.1.: Screenshots of the search and browse interface of the rye *GenomeZipper* available at the PGSB PlantsDB database (<http://mips.helmholtz-muenchen.de/plant/rye/gz/index.jsp>).

4.2. Insights into the structure, organisation, and evolution of cereal genomes

blocks, (ii) to assess genes within the conserved segments, (iii) to detect ancient whole genome duplications, and (iv) to reconstruct the ancestral grass karyotype [159, 345].

The extensive synteny conservation observed in grasses has played a major role in this thesis. It laid the basis for comprehensive comparative analyses between the low-pass shotgun sequenced barley and rye genomes, the wheat deletion bin-mapped ESTs [346], and the fully sequenced genomes of *Brachypodium* [79], rice [75], and sorghum [80]. The information gained from these analyses has proven to be of high value for a series of tasks and topics. Namely it was used

- (i) to generate virtually ordered gene maps for the barley and rye chromosomes,
- (ii) to investigate lineage-specific evolutionary events among grass genomes,
- (iii) to provide a revised model of rye genome evolution and speciation, including a characterisation of its interspecific rearrangements,
- (iv) to characterise the composition of rye B chromosomes, and
- (v) to trace back the origin and evolution of rye B chromosomes.

4.2.1. The barley and rye genomes

Comparisons of barley and rye genomes against the three model grass genomes not only enabled the estimation of gene numbers in the two cereals, but also facilitated the identification of syntenically conserved linkage groups. The identified collinear linkage groups are consistent with previously reported genome collinearity among *Triticeae* and the three reference genomes [79, 319, 347]. With the unprecedented high resolution, analyses give a detailed insight into the degree of synteny conservation among the rye and barley genes. It reveals that 30% to 40% of the rye and barley genes are not allocated in the syntenic conserved gene space, numbers that are also supported by findings in wheat [64, 348, 349]. Despite the large number of rearrangements observed in rye relative to barley, both cereal genomes share a similar amount of conserved syntenic genes with the reference genomes. For example, the linkage group of chromosome 1 has been extensively studied and it has been proven highly collinear among the *Triticeae* species [64, 350]. Both homoeologous rye and barley chromosomes 1R and 1H confirm the well-conserved gene content and order. Chromosomes 1R and 1H show collinearity to the distal regions of both arms of rice chromosome 5 and sorghum chromosome 9, as well as to the proximal regions of *Brachypodium* short and long arms of chromosome 2. Furthermore, blocks of conserved synteny to 1R and 1H have been found for the *Brachypodium* chromosome 3, rice chromosome 10, and sorghum chromosome 1.

For the genes assigned to collinear linkage blocks the degree of similarity has been explored and several duplicated segments have been identified. Therefore, the rye and barley genome zipper models have been compared to each other, and to the wheat EST markers, *Brachypodium*, rice, and sorghum protein-coding genes. The gene orientation within the syntenically conserved chromosome segments of barley and the three reference genomes has revealed numerous local inversions. These inversions could be attributed either to barley, and/or to one of the reference

4. Discussion and conclusions

genomes. The investigation of ancestral duplications in the barley genome has ratified previous findings [351, 352], but also revealed additional local duplications. Nine complex segmental duplications have been detected, whereby six of these belong to the previously described duplications (duplicated blocks: 6H-2H, 6H-7H, 4H-5H, 4H-1H, 4H-2H, and 1H-3H) and three are barley specific (duplicated regions: 3H-7H, 3H-4H, and 2H). Overall, the duplicated segments cover roughly 48% of the entire barley genome [99].

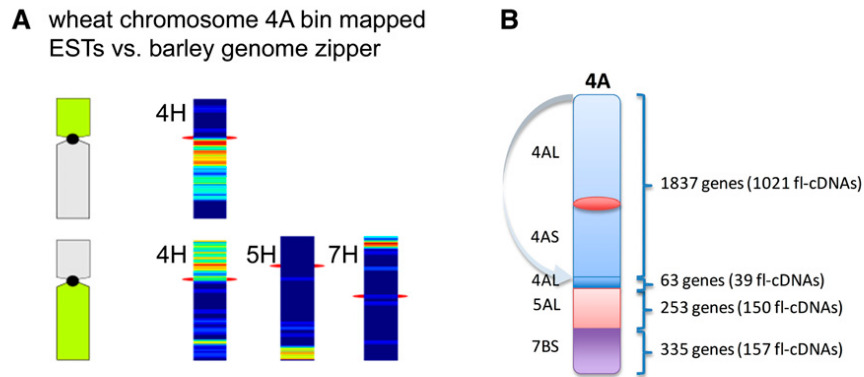


Figure 4.2.: Structure of wheat chromosome 4A in relation to the barley pseudo-chromosomes. The ESTs marker of wheat chromosome 4A have been compared to the virtually ordered barley pseudo-chromosomes. The syntenic conserved regions are depicted and displayed by a heatmap. ‘(A) Wheat EST markers allocated to 4AS cross-match to barley genes on 4HL and markers allocated to 4AS, a small region on 4AL, 5AL, and 7BS cross-match to 4HL. Thus, a reciprocal translocation involving chromosomes 4A and 5A and a translocation from 7BS to 4AL was detected. Compared with barley 4H, wheat chromosome 4A contains a pericentromeric inversion. (B) The barley genome zipper model allows the size of the affected regions to be estimated and the minimal number of genes located in these rearranged regions of the wheat chromosomes to be predicted.’ Figure and legend from [99], by kind permission from American Society of Plant Biologists.

The assembled barley fl-cDNA gene indices and wheat deletion bin-mapped ESTs [346] have been compared to survey the structural variation and gene content between wheat and barley. The comparison results confirm the close relatedness and extensive synteny conservation shared by the genomes of wheat and barley, that has been attested by several studies over the past decades, too [154, 319, 353–356]. Only few structural differences are found among genomes, most of them involving the highly rearranged wheat chromosome 4A (see Fig. 4.2). The observed rearrangements are wheat specific and have occurred after the split of wheat and barley in the common ancestor of the wheat A subgenome. They involve an extensive pericentromeric inversion on 4A and two interchromosomal translocations between the chromosomes 4A and 5A, and between the long chromosome arm of 4A and the short chromosome arm of 7B [162, 357]. The description of these structural variations is well-known and has been discussed in detail [346, 358]. However, the comparison based on the barley genome zipper model exceeds the

4.2. Insights into the structure, organisation, and evolution of cereal genomes

resolution and allows the estimation of the size of the rearranged chromosome fragments, as well as an approximation of the number of genes located within them. In addition to the above mentioned pericentromeric inversion the high-density comparisons conducted in this thesis support the existence of additional pericentric inversions on several wheat chromosomes (2B, 5A, 3B, 6B) [353, 358]. The inversions observed on wheat chromosome group 1 may indicate an inversion that has occurred in barley, but additional evidences are necessary for validation. Recently, Ma et al. [359] have shown that pericentromeric inversions frequently occur in wheat, finding strong evidences not only for the previously reported inversions, but also for other inversions on wheat chromosomes 2D, 4B, 6A, 7A, and 7B. Clearly, with the availability of complete sequenced genomes the detection of further rearrangements remains possible.

The comparison also reveals differences in the conservation pattern between barley and the wheat A, B, and D subgenomes. Small regions appear to be missing from individual wheat subgenomes, namely on chromosome groups 1, 3, and 5. For example, on the long chromosome arm of barley 1H a region has been localised, that is absent on chromosome 1 in all three wheat subgenomes. Unfortunately the available data resolution does not permit to conclude whether this region (i) has been excluded from wheat genome by gene loss, (ii) has been included in the barley genome after the split from the wheat lineage, or (iii) is not represented in the available wheat EST data set. The pattern variation found among homoeologous wheat chromosomes might be pre-existing or might indicate segmental gene loss during or subsequent to the polyploidisation event. This suggests that differences among the wheat A, B, and D subgenomes may be found rather at functional and regulatory level than at a structural level.

The homoeologous relationship between rye and wheat was investigated several decades ago by Devos et al. [161]. Using roughly 150 RFLP low resolution markers, the authors have provided a first insight into the rye genome structure, indicating that the rye genome has undergone multiple translocations relative to that of wheat. In this thesis the conserved synteny among rye and *Triticeae* cereals has been reassessed. Barley was used as a reference and genome zipper models provided an unprecedented resolution for both genomes. The result of this analysis largely confirms earlier observations [160, 161]. Overall, the rye genome can be subdivided into 17 chromosomal segments that share collinearity with the barley genome and which record the evolutionary development of the modern rye genome. The chromosomal rearrangements involve six out of seven rye chromosomes, namely the chromosomes 2R to 7R. These chromosomes can be depicted as a mosaic pattern ranging from two to four segments, which are equivalent to individual regions in barley. Chromosome 1R is an exception. It is lacking any structural variation and is homoeologous to barley chromosome 1H over its entire length. Since the observed rearrangements exhibit a particular pattern, it can be concluded that they have occurred as a succession of six translocation events. Based on these events a revised model of genome evolution has been proposed [102]. Assuming that the genome organisation of the last common ancestor of rye and barley resembled the modern genome of barley, four out of six translocation events could be ordered chronologically. The a4/a5 translocation has been defined as the first step in this series of events, due to its similarity to the reciprocal translocation indicated between the wheat chromosomes 4A and 5A [162, 357]. In the next steps the ancestral chromosomes have been involved in three more subsequent translocations, namely between the chromosomes a3 and a6, a6 and a7, and a7 and a4. The sequential order of the last two translocation events (a2/a7

4. Discussion and conclusions

and a6/a4) could not be specified, but most likely they succeeded the previous events and may have occurred at the same time.

The mosaic structure of the rye genome observed in this thesis raises the question which evolutionary mechanisms have shaped the modern rye genome. To address this question the conservation patterns of orthologous genes from the rye conserved segments and their counterparts in barley have been compared against each other and against the reference genomes of *Brachypodium*, rice, and sorghum. The results indicate major differences in the sequence conservation of the individual rye and barley segments as well as a dissimilar number of conserved orthologous genes with regard to the reference grass genomes. Since no variation in selection pressure could be found between the individual rye segments, phylogenetic analyses were conducted to elucidate the patterns and dynamics of rye speciation. The inferred evolutionary relationships of rye to the genomes of barley, *Triticum monococcum*, *Aegilops tauschii*, *Brachypodium*, rice, and sorghum revealed various incongruent trees, which are consistent with the previously observed differences in sequence similarity among some of the 17 rye fragments. For 8 out of 17 fragments the constructed consensus trees located rye between the barley and wheat lineage, while the other ones exhibit distinct patterns. Hence, the evolutionary history of rye turns out to resemble a network rather than a tree structure. Furthermore, four of the rye fragments indicate evidences of reticulate evolution or introgressive hybridisation, which led to the conclusion that the rye genome can be depicted as an aggregation of genomic segments, which may stem in part from varying evolutionary origins. The phylogenetic incongruences between rye and the other grasses identified in this work also support earlier observations that have been made by Escobar et al. [360]. Therein the authors explored the phylogenetic relationships of several *Triticeae* species, which also include rye, based on only few chloroplast and nuclear genes and hypothesised that rye exhibits signatures of reticulate evolution.

Reticulate or hybrid speciation are common in some fungi species [361] and marine organisms [362–366], but also in flowering plants, like sunflowers, soybean, *Brassica*, and *Triticeae* species [360, 367–375]. It has been shown that it can occur by both diploid and allopolyploid hybrid speciation (reviewed in [376]) and that it may offer selective advantages to the new arisen species under changing environmental conditions or in more extreme habitats [377–379]. This implies that hybridisation is a major evolutionary mechanism for speciation and that it may be the underlying mechanism that shaped the modern rye genome. However, it remains unclear whether rye has been formed by allopolyploid or diploid hybrid speciation, even though both events are conceivable. The observed mosaic chromosome structure may have resulted from one or more polyploidisation and/or interspecific hybridisation events with (yet unknown) related species, followed by subsequent diploidisation and comprehensive chromosome reorganisations. An allopolyploid origin of the rye genome followed by extensive chromosome restructuring would also explain the increased genome size (8.1 Gbp) and gene content of rye in comparison to the genomes of barley and wheat. On the other hand it might have evolved from one or more diploid hybrid speciation events. This assumption is supported by the diploid, outbreeding nature of the rye genome. The hypothesis of rye evolution via introgressive hybridisation (both diploid and polyploid) may explain the observed divergent level of sequence homology to barley and wheat, as well as the numerous reciprocal translocations that restructured the modern rye genome.

4.2.2. Deciphering the rye B chromosome

Apart from the characterisation of the rye and barley genomes, this thesis has also explored the organisation, structure, and evolution of the rye B chromosomes. B chromosomes are dispensable, selfish genetic elements that exist in addition to the standard karyotype of several species and which do not exhibit Mendelian inheritance patterns (see section 1.3). Comparative sequence analysis of a flow-sorted rye B chromosome and the virtual gene map of barley allowed to trace back the origin of rye B chromosomes and to propose an evolutionary model for their origin in rye. The rye B sequence reads could be mapped on five different barley chromosomes (3H, 2H, 4H, 5H, and 7H), suggesting an intraspecific, multichromosomal origin of the rye B chromosome. Taking into account the well-preserved synteny conservation among the rye and barley genomes, as well as the chromosomal rearrangements that the rye genome underwent during its evolution, these regions could be assigned to the rye A chromosomes 3RS and 7R (see Fig. 1 in [295]). The short arm of rye chromosome 3R is mainly collinear with the region on barley chromosome 3H, whereas rye chromosome 7R shows collinearity to the barley chromosomal regions 2H, 4H, 5H, and 7H. The multi-A chromosome origin found in this work is further supported by thousands of genomic segments derived from all rye A chromosomes. Their accumulation has most probably occurred after the formation of a proto-B chromosome from the A chromosomes 3RS and 7R, and may have resulted from (i) double-strand break repair [380], (ii) transposition of genomic sequences, as it has been shown for non-collinear *Triticeae* genes [64], or (iii) sequences eliminated from the A chromosomes during the structural rearrangements that have taken place in rye. A multi-A-chromosome origin has been proposed also for the B chromosomes of maize [215], *Brachycome dichromosomatica* [221], the cichlid fish *Astatotilapia latifasciata* [381], and *Astyanax paranae* [382].

Recently, the virtual gene map of rye has become available [102] and the B chromosome origin proposed in this work has been validated (see Fig. 1 in [383]). In addition, the presence of B chromosomes in the rye genome reinforce the hypothesis that interspecies hybridisation has shaped the rye genome and influenced its evolution. Thus, it can be presumed that the genesis of the rye B chromosomes may be a by-product of the rye genome reorganisation after introgressive hybridisation, or rediploidisation. These events then were followed by one or more whole/segmental genome duplications and sequence insertions. Most likely this initial step was followed by further adaptations, so that the B chromosomes evolved their own evolutionary dynamics. Consequently, the recombination of the proto-B chromosome with its donor chromosomes became restricted and rapid structural modifications might have been initiated to establish an own drive mechanism. Therefore, the coding and non-coding sequences may have been subjected to silencing and degeneration through mutations and sequence insertions as a result of relaxed selective pressure. An exception are those sequences involved in the maintenance of the B chromosomes, namely in their drive mechanism. Lately, it has been shown that the repetitive elements located in the nondisjunction control region of rye Bs are transcriptionally active [262, 384]. Furthermore, this region is highly conserved not only regarding its transcription activity, but also regarding its replication and histone composition among several geographically distinct populations of cultivated (*Secale cereale* ssp. *cereale*) and weedy rye (*Secale cereale* ssp. *segetale*) [299]. The observed conserved molecular structure of the B chromosomes at the level of subspecies also suggests its monophyletic origin [299]. It is conceivable that the B

4. Discussion and conclusions

chromosome evolutionary model proposed above is found mainly in species with an increased incidence of chromosomal rearrangements and phylogenetic groups with unstable chromosome numbers [295].

The most unexpected insight of this thesis is that rye B chromosomes carry large amounts of A-derived genic sequences. The discovery of rye B-located gene fragments contradicts the common assumption that B chromosomes are genetically inert without any functional genes. These gene fragments represent copies of A chromosome genes, showing different levels of sequence similarity to the corresponding A homologous genes. Although no conclusions can be drawn from the present data regarding their completeness and functionality, these findings may contribute to the explanation why the presence of B chromosomes is associated with deleterious effects. It is likely that 'dosage compensation' regulates the coexistence of sequence-identical A- and B-derived transcripts. Hence, B-located genes could affect the transcriptome profile of their host genome and cause considerable damage or confer a fitness advantage to their host. Until lately, most studies focussed on the detection of functional high-copy number genes, neglecting the identification of low-copy genes based on the assumption that Bs lack them. The majority of the B-located genic sequences are highly fragmented and most probably pseudogenes, which reflects their faster degeneration or earlier insertion in Bs [381, 385]. However, some of them were intact and their expression could be confirmed (reviewed in [383]). Recent data suggest that not only rye B chromosomes carry putative transcriptionally active genes [385], but also the B chromosomes of Siberian roe deer [386], of canid [269], of cichlid fish *Astatotilapia latifasciata* [381], of *Astyanax paranae* [382], and of *Drosophila albomicans* [387].

In addition to the numerous gene fragments, rye B chromosomes also accumulate significantly higher amounts of chloroplast and mitochondrial DNA than the A chromosomes. The organellar inserts found on the Bs cover almost the entire chloroplast and mitochondrial genomes, indicating that all sequences are transferable from organelle to nucleus. Since B chromosomes are not necessary for the development and growth of the host species, it is likely that they tolerate a higher mutation frequency and DNA insertions rate. Thus, the interplay of frequent sequence integration and rapid elimination seems to be imbalanced for B chromosomes. The insertions most probably have fewer deleterious genetic consequences relative to those on the A chromosomes, supporting the assumption of a reduced selection on supernumerary chromosomes. It has also been observed that mitochondrial DNA inserts are located around the pericentromeric region of Bs. Marques et al. [299] have examined in greater detail the location of the organellar inserts on the rye B chromosomes. While the chloroplast DNA is located only on the long chromosome arm of rye Bs, the mitochondrial DNA is detected mainly on the pericentromeric region of the long chromosome arm. Furthermore, the authors observed in weedy rye a structural variant of the B chromosome, which differs from the standard one by a pericentric inversion. Apart from this inversion, which seems to have occurred during the evolution of the rye B chromosomes, the analysed distribution of repeat sequences and organellar inserts is highly conserved, despite their different geographic population origin [299]. Lately, Ruban et al. [388] have demonstrated that rye B chromosomes are not the only Bs that accumulate organellar DNA. The B chromosomes of *Aegilops speltoides* concentrate a considerable amount of organellar DNA, too, even if the distribution of the B-located organellar inserts differs among the tested populations [388]. Nevertheless, further analyses are necessary to address the question whether the organellar DNA

4.2. Insights into the structure, organisation, and evolution of cereal genomes

transfer on B chromosomes is an important mechanism that drives the Bs evolution.

The new knowledge gained during this research raises additional questions, such as whether duplicated genes are located on Bs, whether B-located genes are translated to functional proteins, whether B-located genes affect the epigenetic status of A-located genes, and how B-located genes are regulated [383]. These questions cannot be answered completely unless B chromosomes are fully sequenced and annotated. However, the use of B chromosomes for the development of artificial chromosomes seems promising. Their unique features, like dispensability and the lack of meiotic pairing with A chromosomes, and the ability to sort them easily by flow cytometry would allow to make use of them for chromosomal engineering in both mammals and plants [389]. They could facilitate the increase of resistance to diseases, the study of gene dosage effects, transfer genes to the karyotype, and the development of novel approaches for the prevention, diagnosis and treatment of medical conditions. Therefore, further efforts should be made to determine their annotation and transcriptional activity, as well as to comprehend their regulatory mechanisms and their genotype-phenotype correlations.

A. Configuration Files

Listing A.1: Configuration file of the *GenomeZipper* wrapper.

```
##The GenomeZipper pipeline can be run via the script "zipperPipeline_wrapper.pl".
##The script accept a config file as parameter. The config file set the parameter
##for all substeps of the pipeline, which can be run individually or as a combina-
##tion of steps.

##Required values:
MASKING:          #yes/no - run repeat masking to filter out the repeats
SYNTENY:         #yes/no - run ChromoWIZ to detect syntenic regions
ZIPPER:          #yes/no - run the zipper step

WORK_DIR:        #working directory (will be created if it doesn't exist)
QUEUE:           #SGE queue name (e.g. plant2)
Q_NAME:          #query name (e.g. 3H)
Q_FASTA:         #query fasta file
Q_TYPE:          #data type (default: contigs, e.g. contigs, reads, genes, all)

##STEP1: repeat masking
REPEAT_PARAM:
  LIB:           #repeat library name (e.g. REdat9.0)
  VMATCH:        #path to Vmatch
  INDEX:         #path to the repeat library index file
  LEN:           #alignment length (e.g. 100)
  ID:            #identity value (e.g. 70)
  EXDROP:        #exdrop value (e.g. 5)
  SEED:          #seed value (e.g. 14)
  EVAL:          #evalue (e.g. 1e-03)
  DESC:          #description (e.g. 0)
  MASK:          #letter used to mask the repeats (e.g. N)

##STEP2: synteny
SYNTENY_PARAM:
  BLAST_TYPE:    #Blast type (e.g. blastx, blastp, blastn)
  WIN_SIZE:      #windows size in base pairs (e.g. 500000)
  SHIFT:         #shift size in base pairs (e.g. 100000)
  MIN_CHR:       #size (bp) of the smallest chromosome (e.g. 20000000)
  ANNO:          #"_cds"
  GFF_TYPE:      #gff type (e.g. "CDS,mRNA")
  GAP:           #gap size in percent (e.g. 20)
  INTENSITY:     #colour intensity (e.g. 40)
  REF1:
    NAME:        #shortcut first ref. genome (e.g. Bd)
    SIM:         #blast similarity (e.g: 75), default = mean identity
    ALLEN:       #alignment length (bp or aa, e.g. 30 aa or 100 bp)
```

A. Configuration Files

```
REF1_FA:           #CDS or protein fasta
REF1_GENOME:       #genomic sequence
REF1_GFF:          #gff3 file
REF2:
  NAME:            #shortcut second ref. genome (e.g. Os)
  SIM:             #blast similarity (e.g. 70), default = mean identity
  ALN:            #alignment length (bp or aa, e.g. 30 aa or 100 bp)
  REF2_FA:         #CDS or protein fasta
  REF2_GENOME:     #genomic sequence
  REF2_GFF:        #gff3 file
REF3:
  NAME:            #shortcut third ref. genome (e.g. Sb)
  SIM:             #blast similarity (e.g. 70), default = mean identity
  ALN:            #alignment length (bp or aa, e.g. 30 aa or 100 bp)
  REF3_FA:         #CDS or protein fasta
  REF3_GENOME:     #genomic sequence
  REF3_GFF:        #gff3 file

##STEP3: GenomeZipper
ZIPPER_PARAM:
  DIR:             #output directory (will be created if it doesn't exist)
  SLICE:           #nr. of elements to regard during the sorting step (default 11)
  MARKER_ID:       #tab-delimited file with marker name and cM position
  MARKER_FA:       #marker fasta
  FLCDNA_ID:       #tab-delimited file with fl-cDNAs name and chromosome
  FLCDNA_FA:       #fl-cDNAs fasta
  EST_FA:          #EST fasta
  REF1_NAME:       #shortcut first ref. genome
  REF1_NUC_FA:     #CDS fasta for the first ref. genome
  REF2_NAME:       #shortcut second ref. genome
  REF2_NUC_FA:     #CDS fasta for the second ref. genome
  REF3_NAME:       #shortcut third ref. genome
  REF3_NUC_FA:     #CDS fasta for the third ref. genome
  Q_MARKER_SIM:    #similarity value for query marker matches
  Q_EST_SIM:       #similarity value for query EST matches
  Q_FL_SIM:        #similarity value for query fl-cDNAs matches
  Q_REF1_SIM:      #similarity value for query reference1 matches
  Q_REF2_SIM:      #similarity value for query reference2 matches
  Q_REF3_SIM:      #similarity value for query reference3 matches
  #THE FOLLOWING PARAMETER ARE SET ONLY IF THE SYNTENY STEP IS SET ON "NO"
  TAG_GEN1:        #all tagged genes from ref. genome 1 (tab-delimited)
  TAG_GEN2:        #all tagged genes from ref. genome 2 (tab-delimited)
  TAG_GEN3:        #all tagged genes from ref. genome 3 (tab-delimited)
  SELECTION1:      #name tagged gene from ref1 in syn. region + chromosome
  SELECTION2:      #name tagged gene from ref2 in syn. region + chromosome
  SELECTION3:      #name tagged gene from ref3 in syn. region + chromosome
  BBH_MARKER_REF1: #BBH pairs (tab-delimited: marker gene_ref1)
  BBH_MARKER_REF2: #BBH pairs (tab-delimited: marker gene_ref2)
  BBH_MARKER_REF3: #BBH pairs (tab-delimited: marker gene_ref3)
  BBH_MARKER_EST:  #BBH pairs (tab-delimited: marker ESTs)
  BBH_MARKER_FL:   #BBH pairs (tab-delimited: marker fl-cDNAs)
  BBH_REF1_REF2:   #BBH pairs (tab-delimited: gene_ref1 gene_ref2)
  BBH_REF1_REF3:   #BBH pairs (tab-delimited: gene_ref1 gene_ref3)
  BBH_REF2_REF3:   #BBH pairs (tab-delimited: gene_ref2 gene_ref3)
```

```

BBH_REF1_FL:      #BBH pairs (tab-delimited: gene_ref1 fl-cDNAs)
BBH_REF2_FL:      #BBH pairs (tab-delimited: gene_ref2 fl-cDNAs)
BBH_REF3_FL:      #BBH pairs (tab-delimited: gene_ref3 fl-cDNAs)

```

Listing A.2: Configuration file for the calculation of the syntenic regions

```

##Use this configuration file to run the synteny step. The file is
##automatically generated by the GenomeZipper wrapper. The configu-
##ration file depicted here can be used to compute all syntenic
##regions to three individual reference genomes.

WORK_DIR:          #working directory
CHRID_1:
  ANNO_ID:          #query name (e.g. 3H)
  TYPE:             #data type (e.g. contigs)
  GENOMES:
    FIRST:
      NAME:          #shortcut first reference genome (e.g. Bd)
      BLAST_TYPE:    #path to blast output (query vs. ref. genome)
      TAGGED_GENES: #tab-delimited output file with all tagged ref. genes
      TAGGED_GFF:    #gff output file with all tagged ref. genes
      SELECTION:     #tab-delimited output file with all syntenic ref. genes
      LEN:           #alignment length in aa (e.g. 30)
      ID:            #similarity value (e.g. 75)
      CHROMO_WIZ:
        WIN_SIZE:    #windows size (e.g. 500000)
        SHIFT:       #shift size (e.g. 100000)
        MIN_CHR:     #mi\~ni\~mal chromosome length (e.g. 20000000)
        ANNO:        "_cds"
        SQL:         #path to sqlliteDB
        SEQ_REF:     #genome sequence of the ref. genome
        GFF_REF:     #corresponding gff3 file
        RUN_GFF3:    "yes"
        RUN_DENS:    "yes"
        GFF_TYPE:    #gff type (e.g. "CDS,mRNA")
        GAP:         #gap size in percent (e.g. 20)
        INTENSITY:   #colour intensity (e.g. 40)
        OUT:         #output directory
    SECOND:
      NAME:          #shortcut first reference genome (e.g. Os)
      BLAST_TYPE:    #path to blast output (query vs. ref. genome)
      TAGGED_GENES: #tab-delimited output file with all tagged ref. genes
      TAGGED_GFF:    #gff output file with all tagged ref. genes
      SELECTION:     #tab-delimited output file with all syntenic ref. genes
      LEN:           #alignment length in aa (e.g. 30)
      ID:            #similarity value (e.g. 70)
      CHROMO_WIZ:
        WIN_SIZE:    #windows size (e.g. 500000)
        SHIFT:       #shift size (e.g. 100000)
        MIN_CHR:     #mi\~ni\~mal chromosome length (e.g. 20000000)
        ANNO:        "_cds"
        SQL:         #path to sqlliteDB
        SEQ_REF:     #genome sequence of the ref. genome
        GFF_REF:     #corresponding gff3 file

```

A. Configuration Files

```
RUN_GFF3: "yes"
RUN_DENS: "yes"
GFF_TYPE: #gff type (e.g. "CDS,mRNA")
GAP:      #gap size in percent (e.g. 20)
INTENSITY: #colour intensity (e.g. 40)
OUT:      #output directory
THIRD:
NAME:     #shortcut first reference genome (e.g. Sb)
BLAST_TYPE: #path to blast output (query vs. ref. genome)
TAGGED_GENES: #tab-delimited output file with all tagged ref. genes
TAGGED_GFF: #gff output file with all tagged ref. genes
SELECTION: #tab-delimited output file with all syntenic ref. genes
LEN:      #alignment length in aa (e.g. 30)
ID:       #similarity value (e.g. 70)
CHROMO_WIZ:
WIN_SIZE: #windows size (e.g. 500000)
SHIFT:    #shift size (e.g. 100000)
MIN_CHR:  #mi\~ni\~mal chromosome length (e.g. 20000000)
ANNO:     "_cds"
SQL:      #path to sqlliteDB
SEQ_REF:  #genome sequence of the ref. genome
GFF_REF:  #corresponding gff3 file
RUN_GFF3: "yes"
RUN_DENS: "yes"
GFF_TYPE: #gff type (e.g. "CDS,mRNA")
GAP:      #gap size in percent (e.g. 20)
INTENSITY: #colour intensity (e.g. 40)
OUT:      #output directory
```

Listing A.3: *GenomeZipper* configuration file.

```
##Use this configuration file to run the zipper step. The file is
##automatically generated by the GenomeZipper wrapper. The required
##Blasts and BBH can be run manually or automatically by the Genome-
##Zipper wrapper.

QUERY:      #query name (e.g. 3H)
TYPE:       #data type (e.g. contigs)
SLICE:      #nr. of elements to regard during the sorting step (default 11)

MARKER_IDS: #tab-delimited file with marker name and cM position
QUERY_IDS:  #tab-delimited file with query name and chromosome
FL_IDS:     #tab-delimited file with fl-cDNAs name and chromosome

OUT_TAB:    #virtually ordered gene map in tab-delimited format
OUT_XLS:    #virtually ordered gene map in Excel format
OUT_CSV:    #virtually ordered gene map in csv-delimited format
OUT_STAT:   #output file with GenomeZipper statistics

NR_REF: 3
REF1:
NAME:      #shortcut first ref. genome (e.g. Bd)
SYN:       #list with syntenic ref. genes and position
ORTHO:     #list with all tagged ref. genes
```


REF2:
NAME: #shortcut first ref. genome (e.g. Os)
SYN: #list with syntenic ref. genes and position
ORTHO: #list with all tagged ref. genes
REF3:
NAME: #shortcut first ref. genome (e.g. Sb)
SYN: #list with syntenic ref. genes and position
ORTHO: #list with all tagged ref. genes

QUERY_MARKER: #first best hit (tab-delimited: query marker)
QUERY_EST: #first best hit (tab-delimited: query ESTs)
QUERY_FL: #first best hit (tab-delimited: query fl-cDNAs)

BBH_REF1_FL: #BBH pairs (tab-delimited: gene_ref1 fl-cDNAs)
BBH_REF2_FL: #BBH pairs (tab-delimited: gene_ref2 fl-cDNAs)
BBH_REF3_FL: #BBH pairs (tab-delimited: gene_ref3 fl-cDNAs)

BBH_MARKER_REF1: #BBH pairs (tab-delimited: marker gene_ref1)
BBH_MARKER_REF2: #BBH pairs (tab-delimited: marker gene_ref2)
BBH_MARKER_REF3: #BBH pairs (tab-delimited: marker gene_ref3)
BBH_MARKER_EST: #BBH pairs (tab-delimited: marker ESTs)
BBH_MARKER_FL: #BBH pairs (tab-delimited: marker fl-cDNAs)
BBH_REF1_REF2: #BBH pairs (tab-delimited: gene_ref1 gene_ref2)
BBH_REF1_REF3: #BBH pairs (tab-delimited: gene_ref1 gene_ref3)
BBH_REF2_REF3: #BBH pairs (tab-delimited: gene_ref2 gene_ref3)

BBH_QUERY_REF1: #BBH pairs (tab-delimited: query gene_ref1)
BBH_QUERY_REF2: #BBH pairs (tab-delimited: query gene_ref2)
BBH_QUERY_REF3: #BBH pairs (tab-delimited: query gene_ref3)
BBH_QUERY_EST: #BBH pairs (tab-delimited: query ESTs)
BBH_QUERY_FL: #BBH pairs (tab-delimited: query fl-cDNAs)
BBH_QUERY_MARKER: #BBH pairs (tab-delimited: query marker)

B. Original publications

This appendix comprise the four original articles presented and discussed in this thesis. The articles are reprinted with permission of the respective publisher:

1. K.F.X. Mayer, S. Taudien, **M. Martis**, H. Šimková, P. Suchánková, H. Gundlach, T. Wicker, A. Petzold, M. Felder, B. Steuernagel, U. Scholz, A. Graner, M. Platzer, J. Doležel, and N. Stein. *Gene Content and Virtual Gene Order of Barley Chromosome 1H*. **Plant Physiol** 2009, 151: 496–505, doi:10.1104/pp.109.142612.
Downloaded from www.plantphysiol.org
©2009 American Society of Plant Biologists
2. K.F.X. Mayer, **M. Martis**, P.E. Hedley, H. Šimková, H. Liu, J.A. Morris, B. Steuernagel, S. Taudien, S. Roessner, H. Gundlach, M. Kubaláková, P. Suchánková, F. Murat, M. Felder, T. Nussbaumer, A. Graner, J. Salse, T. Endo, H. Sakai, T. Tanaka, T. Itoh, K. Sato, M. Platzer, T. Matsumoto, U. Scholz, J. Doležel, R. Waugh, and N. Stein. *Unlocking the Barley Genome by Chromosomal and Comparative Genomics*. **Plant Cell** 2011, 23: 1249–1263, doi:10.1105/tpc.110.082537.
Downloaded from www.plantcell.org
©2011 American Society of Plant Biologists
3. **M.M. Martis**, S. Klemme, A.M. Banei-Moghaddam, F.R. Blattner, J. Macas, T. Schmutzer, U. Scholz, H. Gundlach, T. Wicker, H. Šimková, P. Novák, P. Neumann, M. Kubaláková, E. Bauer, G. Haseneyer, J. Fuchs, J. Doležel, N. Stein, K.F.X. Mayer, and A. Houben. *Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences*. **PNAS** 2012, 109(33): 13343–13346, doi:10.1073/pnas.1204237109.
Downloaded from www.pnas.org
4. **M.M. Martis**, R. Zhou, G. Haseneyer, T. Schmutzer, J. Vrána, M. Kubaláková, S. König, K.G. Kugler, U. Scholz, B. Hackauf, V. Korzun, C.C. Schön, J. Doležel, E. Bauer, K.F.X. Mayer, and N. Stein. *Reticulate evolution of the rye genome*. **Plant Cell** 2013, 25:3585–3698, doi:10.1105/tpc.113.114553.
Downloaded from www.plantcell.org
©2013 American Society of Plant Biologists

Gene Content and Virtual Gene Order of Barley Chromosome 1H¹[C][W][OA]

Klaus F.X. Mayer, Stefan Taudien, Mihaela Martis, Hana Šimková, Pavla Suchánková, Heidrun Gundlach, Thomas Wicker, Andreas Petzold, Marius Felder, Burkhard Steuernagel, Uwe Scholz, Andreas Graner, Matthias Platzer, Jaroslav Doležel, and Nils Stein*

Munich Information Center for Protein Sequences/Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum Munich, German Research Center for Environmental Health, 85764 Neuherberg, Germany (K.F.X.M., M.M., H.G.); Leibniz Institute for Age Research, Fritz Lipmann Institute, 07745 Jena, Germany (S.T., A.P., M.F., M.P.); Laboratory of Molecular Cytogenetics and Cytometry, Institute of Experimental Botany, 77200 Olomouc, Czech Republic (H.Š., P.S., J.D.); Institute of Plant Biology, University of Zurich, CH-8008 Zurich, Switzerland (T.W.); and Leibniz Institute of Plant Genetics and Crop Plant Research, 06466 Gatersleben, Germany (B.S., U.S., A.G., N.S.)

Chromosome 1H (approximately 622 Mb) of barley (*Hordeum vulgare*) was isolated by flow sorting and shotgun sequenced by GSFLX pyrosequencing to 1.3-fold coverage. Fluorescence in situ hybridization and stringent sequence comparison against genetically mapped barley genes revealed 95% purity of the sorted chromosome 1H fraction. Sequence comparison against the reference genomes of rice (*Oryza sativa*) and sorghum (*Sorghum bicolor*) and against wheat (*Triticum aestivum*) and barley expressed sequence tag datasets led to the estimation of 4,600 to 5,800 genes on chromosome 1H, and 38,000 to 48,000 genes in the whole barley genome. Conserved gene content between chromosome 1H and known syntenic regions of rice chromosomes 5 and 10, and of sorghum chromosomes 1 and 9 was detected on a per gene resolution. Informed by the syntenic relationships between the two reference genomes, genic barley sequence reads were integrated and ordered to deduce a virtual gene map of barley chromosome 1H. We demonstrate that synteny-based analysis of low-pass shotgun sequenced flow-sorted Triticeae chromosomes can deliver linearly ordered high-resolution gene inventories of individual chromosomes, which complement extensive Triticeae expressed sequence tag datasets. Thus, integration of genomic, transcriptomic, and synteny-derived information represents a major step toward developing reference sequences of chromosomes and complete genomes of the most important plant tribe for mankind.

Access to the complete genome sequence of an organism provides a direct path to gene identification, understanding gene function, exploring genetic diversity, and correlating this information to phenotypic traits. Application of next generation sequencing (NGS) technology (Shendure and Ji, 2008) for whole

genome resequencing may soon become a routine for genome-scale genotyping and haplotype analysis in man. However, such progress is only possible due to the availability of a high-quality reference whole genome sequence—a resource that is still lacking for many of the most important crop species, including the major cereals of the Triticeae tribe.

Barley (*Hordeum vulgare*) is the number four cereal crop in the world. It is a major resource for animal feed and for the brewing and distilling industry. The genome of barley comprises 5.1 Gbp/1 C (Doležel et al., 1998), is about 12 times the size of the rice (*Oryza sativa*) genome, and includes over 80% of repetitive DNA (Schulte et al., 2009; Wicker et al., 2009). The size, high repeat content, and costs of conventional Sanger sequencing impede whole genome sequencing in barley. Consequently, only limited knowledge of its genomic sequence has been accumulated so far by dedicated sequencing of barley bacterial artificial chromosome (BAC) contigs in the course of map-based gene isolation (Stein, 2007). Massive data generation and cost efficiency of NGS allows questions on barley genome composition with unprecedented resolution and depth to be addressed. Wicker et al. (2006, 2009) employed pyrosequencing (454/Roche GS20) to

¹ This work was supported by the program Genome Analysis of the Plant Biological System (www.gabi.de) and by grants from the German Ministry of Education and Research (grant no. BMBF FKZ0314000 to N.S., M.P., K.F.X.M., and U.S.). J.D., H.Š., and P.S. were supported by the Czech Republic Ministry of Education, Youth and Sports (grant no. LC06004). N.S., J.D., K.F.X.M., and T.W. participated within the framework of the European Cooperation in Science and Technology program FA0604.

* Corresponding author; e-mail stein@ipk-gatersleben.de.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Nils Stein (stein@ipk-gatersleben.de).

[C] Some figures in this article are displayed in color online but in black and white in the print edition.

[W] The online version of this article contains Web-only data.

[OA] Open access articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.109.142612

survey gene information on selected barley BAC clones (Wicker et al., 2006) and to catalog the composition of the barley genome (Wicker et al., 2009). Moreover, the short-read sequencing by synthesis (Solexa/Illumina GA 1) was used to generate whole genome shotgun sequence information to assist the statistical annotation of DNA motif frequency at whole genome scale for barley (Wicker et al., 2008). Despite the impressive progress, ordering the massive numbers of short reads obtained by NGS to generate genomic scaffolds of the huge Triticeae genomes remains a major challenge.

Instead of sequencing complex cereal genomes containing large fractions of repetitive DNA, smaller genomes of grass species like rice (1 C to approximately 400 Mbp) and *Brachypodium distachyon* (1 C to approximately 280 Mbp) were suggested as surrogates and models for molecular genomics and positional cloning in cereals with large genomes (Bennetzen and Freeling, 1993; Draper et al., 2001). This strategy is supported by a significant level of colinearity between Poaceae genomes (Moore et al., 1995; Bolot et al., 2009). Moreover, high-quality reference genome sequences for both rice and sorghum (*Sorghum bicolor*) are available (Sasaki and Sederoff, 2003; Paterson et al., 2009) and provide a platform for large-scale implementation of this approach. Although reference genomes represent very important resources of information for molecular genomics in the Triticeae the potential impact of genome colinearity still is limited and can compromise synteny-based gene isolation, since only 50% of the barley genes remain collinear compared to rice (Gaut, 2002; Stein et al., 2007). This observation has been illustrated during map-based cloning of important genes in wheat (*Triticum aestivum*; *vrn2*; Yan et al., 2004) and barley (*vrs1*; Komatsuda et al., 2007) where orthologs were lacking in rice within otherwise well-preserved collinear genome segments.

An additional option to cope with the complexity of cereal genomes is to isolate individual chromosomes and sequence these individually. The reduced complexity of the sorted chromosome samples facilitates molecular analyses, including the isolation of markers and physical mapping (Doležel et al., 2007). Recently, a physical map of wheat chromosome 3B was constructed based on a BAC library cloned from flow-sorted chromosomes (Paux et al., 2006). A procedure for representative amplification of DNA by multiple displacement amplification (MDA) from sorted barley chromosomes was developed (Simkova et al., 2008). As chromosomal DNA in amounts of a few nanograms can be produced easily, this advance opens new avenues for the wider use of chromosome sorting in Triticeae genomics.

In this study, we demonstrate the potential of high-throughput NGS of flow-sorted chromosomes for genome analysis, sequencing, and the development of a high-resolution gene map. As few as 10,000 copies of chromosome 1H were flow sorted from barley cv Morex and used as a template to assess gene content

and genomic composition of this chromosome. Information about sequence conservation and conserved gene content to the rice and sorghum genomes was obtained at unprecedented density and resolution and allowed synteny and homology information to be integrated into a virtual high-density gene map of barley chromosome 1H.

RESULTS

Flow Cytometric Sorting and 454 Sequencing of Barley Chromosomes

Barley has seven chromosomes that are named 1H through 7H according to their homologous relationship to other Triticeae linkage groups (Linde-Laursen, 1996). Flow-cytometric analysis of chromosome suspensions prepared from Morex resulted in histograms of relative fluorescence intensities (flow karyotypes) with a composite peak representing chromosomes 2H to 7H and a small peak of chromosome 1H (Fig. 1). Chromosome 1H is considerably smaller than chromosomes 2H to 7H and can be easily sorted. The sorted fractions of 1H consisted mainly of chromosome 1H ($95.5\% \pm 0.7\%$; mean \pm SD) as determined by fluorescence in situ hybridization (FISH) on 1,000 sorted chromosomes taken during each sort run (data not shown). The contamination was due to various chromosomes and chromosome fragments. Altogether, five batches of 10,000 chromosomes 1H and five batches of 20,000 chromosomes 1H to 7H were prepared for DNA amplification. The amounts of purified DNA

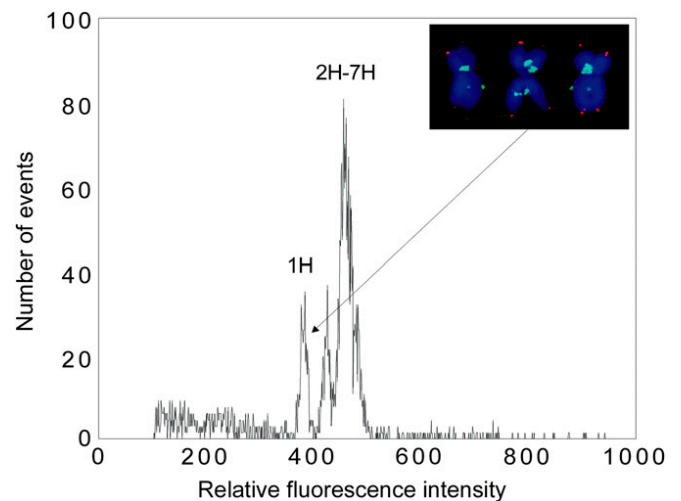


Figure 1. Histogram of fluorescence intensity (flow karyotype) obtained after flow-cytometric analysis of 4',6-diamino-phenylindole stained chromosomes of Morex. The peak of chromosome 1H is well discriminated from the remaining chromosomes forming a composite peak. The insert shows three examples of sorted chromosome 1H after fluorescent labeling of GAA microsatellites (yellow green) and a telomeric repeat (red) using FISH. [See online article for color version of this figure.]

recovered from the sorted chromosomes ranged from 7 to 10 ng and from 10 to 18 ng for chromosome 1H and all chromosomes, respectively. The quantity of DNA obtained after MDA ranged from 3.0 to 5.0 μg DNA in samples with chromosome 1H (whole chromosome amplified 1H = WCA1H), and from 4.5 to 5.6 μg DNA in samples with all chromosomes (1H–7H; whole chromosome amplified all = WCAall).

Enrichment of Chromosome 1H Genomic Sequences

Over 3 million sequence reads comprising close to 800 Mb of sequence were obtained from the shotgun sequence of the flow-sorted chromosome 1H (WCA1H; Table I). Considering the 1 C genome size of barley, 5.1 Gb (Doležel et al., 1998), and relative size of chromosome 1H (12.2%; Marthe and Künzel, 1994), the molecular size of 1H can be estimated to be 622 Mb. Assuming a random distribution of sequence reads, every 200 bp a sequence tag is expected. According to the Lander-Waterman model (Lander and Waterman, 1988) at a 1.29-fold sequence coverage, 72.3% of bases from barley chromosome 1H should be represented in the chromosome shotgun sequence dataset.

We verified the purity in the sorted 1H fractions by comparing the repeat-masked sequence collections from WCA1H to a barley consensus transcript map comprising 2,785 nonredundant EST markers. Chromosome 1H contributed 11.9% (332 markers) of all markers in this map, similar to the relative DNA contribution of chromosome 1H to the entire barley genome (Table II). For the WCA1H sequences, matches were detected to 423 markers of the genome-wide set. A total of 297 out of 332 (89.5%) chromosome 1H located markers were detected whereas only 126 of 2,453 (5.1%) chromosome 2H to 7H markers were hit (cross tab test P value = 0). For sequence data derived from pooled, sorted chromosomes 1H to 7H (WCAall) an even marker detection rate distributed over all chromosomes was observed (Table II). Therefore, based on marker detection rate ($89.5\%/5.1\% = 17.54\%$) and relative contribution of chromosome 1H to the entire barley genome ($87.8\%/12.2\% = 7.2\%$), a 126-fold enrichment ($17.54\% \times 7.2\%$) was observed for WCA1H. This trend was substantiated when using the absolute sequence read counts associated to anchored marker sequences. Of 2,138 individual WCA1H sequence reads anchored to transcript markers, 1,932 (90.4%) were associated with the 297 chromosome 1H markers (Table

II; Fig. 2A). Markers located on chromosomes 2H to 7H accumulated less-frequent WCA1H sequence read matches. One-hundred fifteen of all 126 identified 2H to 7H markers (91%) were hit by three or less WCA1H reads (Table II; Fig. 2B).

We calculated the proportion of detected and undetected markers (true/false positives and negatives, respectively) that were identified (true positives: 297; false positives: 126; true negatives: 2,327; false negatives: 35). A recall rate (sensitivity) of 0.895 and specificity of 0.95 was reached. Applying a confusion matrix, the probability for correct classification reached 0.942. These findings were consistent with the estimated purity of enrichment of 95% estimated by microscopic observation of sorted fractions. In summary, cytological as well as molecular evidence based on marker to sequence read association indicated a 95% purity of the barley WCA1H sequence collection. In addition, the sensitivity exceeded the theoretical expectation of 72% derived from the Lander-Waterman model, as 89.5% of the markers located on chromosome 1 were sequence tagged.

Repeat Composition of the Barley Genome and Chromosome 1H

WCA1H and WCAall datasets were compared for content and frequency of individual classes of repeats. Overall similar fractions of 77.5% (WCA1H) and 74.5% (WCAall) were assigned as repetitive elements. For both datasets, the ratio of class I to class II elements was determined to be 11:1 to 12:1 (Table III). The overall frequency of most element types was very similar; however, deviations were detected for class I retroelements contributing a slightly higher percentage to WCA1H (71.1% versus 67.6% in WCAall). In addition, deviations between datasets were found for CACTA-type elements (6% in WCA1H versus 6.4% in WCAall). The relative amount of ribosomal gene sequences was lower in WCA1H (0.04% versus 0.13% in WCAall). This was consistent with the localization of nucleolus organizing regions on barley chromosomes 6H and 7H (Singh and Tsuchiya, 1982), which thus represent regions that should be depleted in WCA1H.

Estimation of Barley Chromosome 1H Gene Content

To estimate the gene content of chromosome 1H, homology of WCA1H sequence reads to known genes

Table I. 454 sequence read characteristics

Summary of the sequence read characteristics obtained by 454 sequencing of pooled barley chromosomes 1H to 7H (WCAall) and chromosome 1H exclusively (WCA1H).

Dataset	No. Reads Sequenced Before Masking	Total Basepairs	No. Reads After Masking (%)	Median Read Length	M50 Length	N's	No. Reads with Unique Sequences	Reads with Unique Sequences	No. Percent >90 Masked	Masked RepeatMasker
					<i>bp</i>	<i>%</i>		<i>%</i>		<i>bp %</i>
WCAall	381,617	99,401,554	118,779 (31.1%)	256	259	1.96	94,889	79.9	54.3	74.4
WCA1H	3,046,327	799,343,261	896,421 (29.4%)	258	260	2.44	813,914	90.8	56.5	77.5

Table II. 454 read distribution to barley EST-based markers

Sequence reads from pooled chromosomes 1H to 7H (WCAall) and from a chromosome 1H amplified sequence library were compared with 2,785 unique sequence markers anchored on the genetic map of barley. While for WCAall, an even recovery rate over all seven chromosomes was observed, WCA1H is strongly biased toward chromosome 1H. A total of 89.5% of markers recovered and 90.4% of reads are associated with chromosome 1H.

Chromosome	No. of markers	WCAall			WCA1H			Anchored Reads
		No. Reads	No. Marker Detected	Markers Detected	No. Reads	No. Marker Detected	Markers Detected	
				%				%
1H	332	37	26	7.8	1,932	297	89.5	90.4
2H	468	41	32	6.8	38	18	3.8	1.8
3H	445	45	35	7.9	22	13	2.9	1.0
4H	314	32	30	9.6	16	13	4.1	0.7
5H	492	43	36	7.3	57	28	5.7	2.7
6H	337	19	17	5.0	42	29	8.6	2.0
7H	397	28	22	5.5	31	25	6.3	1.4
Total	2,785	245	198		2,138	423		

was surveyed by similarity searches against complete reference genomes, namely rice and sorghum, as well as against clustered EST collections from wheat and barley under optimized stringency conditions (Supplemental Fig. S1, A and B). A total of 4,125 and 4,359 homologous rice and sorghum genes were hit, respectively (BLASTX $\geq 70\%$ identity ≥ 30 amino acids). From wheat and barley EST collections 5,498 and 4,765 (BLASTN) and 3,923 and 4,154 (BLASTX) ESTs and EST clusters were tagged, respectively (Supplemental Table S1). From the comparisons to the different individual reference datasets a nonredundant gene count was extracted comprising 5,126 genes (TBLASTX; $\geq 70\%$ and ≥ 30 amino acids). Given the experimentally observed marker detection rate of approximately 89.5% within the WCA1H dataset, a chromosome 1H content of between 4,600 and 5,800 genes can be estimated. Considering the relative size of chromosome 1H (12%), a total of 38,000 to 48,000 genes can be predicted for the entire barley genome.

Assessment of Conserved Gene Content of Barley Chromosome 1H against Rice and Sorghum

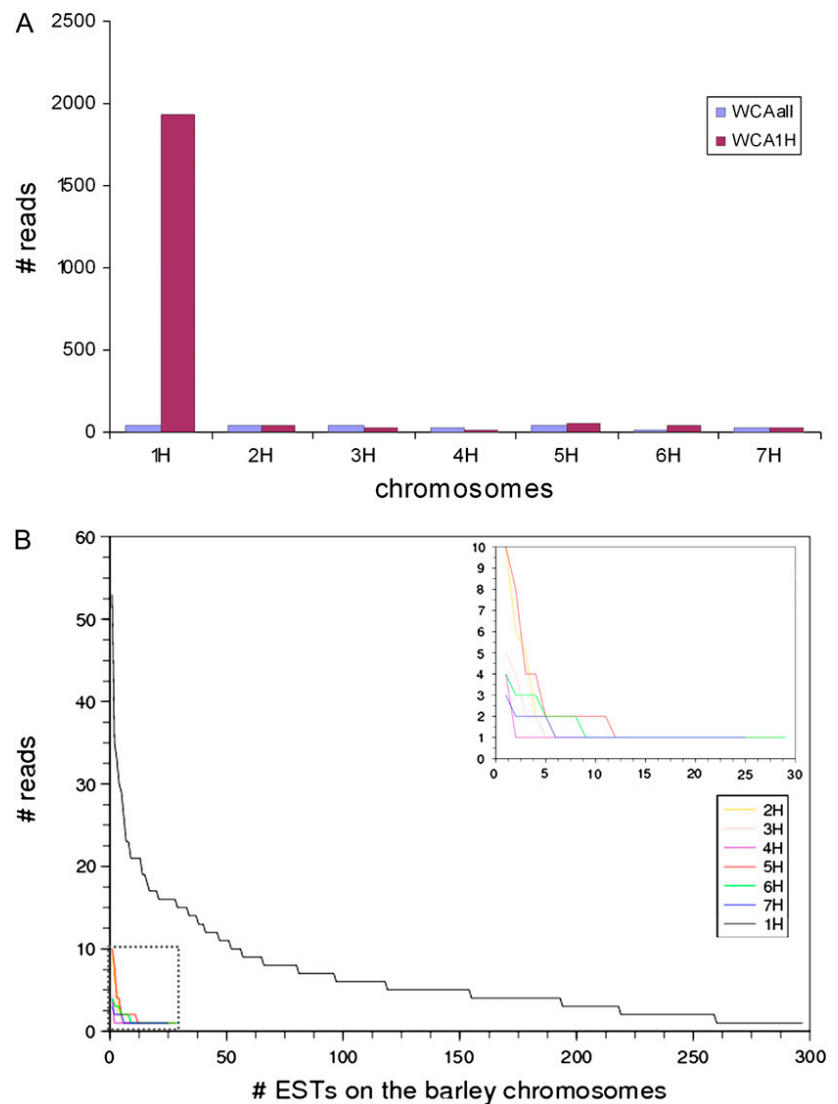
Close syntenic relationships among Poaceae have been known for a long time (Moore et al., 1995). However, the availability of highly enriched chromosome 1H sequence permitted us to infer synteny to rice and sorghum reference genome sequences at the whole chromosome level with a per gene resolution. Using a stringent filter criterion of ≥ 30 amino acid similarity we analyzed the barley WCA1H sequence reads against the respective rice and sorghum genome assemblies and selected for the best homologs. A similar number and percental range of 4,125 (15.2% of all rice genes) and 4,359 (16% of all sorghum genes) homologous genes were detected, respectively (Supplemental Table S3). Rice chromosomes 5 and 10 as well as sorghum chromosomes 1 and 9 were substantially enriched for putative orthologs and outnumbered the remaining chromosomes of the respective

genomes. However, the numbers of putative orthologs provided only a global overview. Therefore, the analysis was refined on the basis of rice and sorghum synteny. Positional information on the respective chromosome was considered and regions containing a high proportion of putative orthologs were depicted (Fig. 3, A and B). Regions with conserved gene content of barley chromosome 1H corresponded to distal regions of both arms of rice chromosome 5 and the distal region of the long arm of rice chromosome 10, respectively. The comparison against sorghum detected such regions for the distal parts of chromosome 9 and the central portion of chromosome 1. A small region of rice chromosome 1 also showed a signal in this analysis. However, subsequent analysis revealed that this region contained a high proportion of protein kinases (26 out of 41 genes) and no apparent synteny to sorghum (data not shown). Generally, genes containing a protein kinase domain are abundant in plant genomes and sequence conservation in the protein kinase domain is usually very high. Therefore, the accumulation of positive matches in this region of rice chromosome 1 indicated rather a false-positive than a true and previously unobserved syntenic region. Due to a lack of detectable syntenic relationship to sorghum and the barley marker scaffold we excluded this region from the subsequent integrative analysis (see below).

Reverse Engineering of an Ordered Gene Map of Barley Chromosome 1H

On the basis of the shotgun read coverage of chromosome 1H, we constructed a virtual gene map of barley chromosome 1H (Fig. 4). Genes from syntenic regions of the rice and sorghum genomes were selected by association with WCA1H sequence reads and were subsequently ordered along the virtual barley chromosome 1H. One hundred and eighty rice and 195 sorghum genes of the syntenic regions could be directly associated to putatively orthologous genetic markers on barley 1H. Their linear order and

Figure 2. Detection of gene-based markers by random (WCAall) and chromosome 1H (WCA1H) sequence collections. A, The number of sequence reads of WCAall and WCA1H samples that could be associated to chromosome-anchored sequence markers was plotted. Sequence reads from the WCAall collection were equally distributed over markers anchored to all seven chromosomes while WCA1H reads were highly enriched for chromosome 1H markers. B, The frequency of WCA1H sequence reads obtained for chromosome 1H compared to 2H to 7H gene-based barley markers differed significantly, respectively. The x axis denotes markers anchored on barley chromosomes 1H to 7H, respectively. The y axis plots the number and distribution of WCA1H sequence reads as observed for markers anchored to individual chromosomes (colored lines). The inset depicts values observed for 2H to 7H. [See online article for color version of this figure.]



synteny association provided the framework for integration and deduction of a virtual gene map of barley chromosome 1H. Out of 1,513 and 1,711 genes contained within the 1H syntenic regions of rice and sorghum, WCA1H sequence reads could be assigned to 1,377 (91%) and 1,551 (90.6%) genes, respectively (Supplemental Table S2). Only these rice and sorghum genes were considered for integration into the virtual barley chromosome 1H gene map (Supplemental Table S4). This approach resulted in tentative anchoring of WCA1H derived sequence tags that detected close to 2,000 putatively orthologous genes from rice and sorghum. Best bidirectional hits revealed orthology between rice and sorghum for 1,174 (1,129 with associated marker or read evidence) genes present in the selected syntenic regions from sorghum and rice. In contrast, 277 (18.31%) rice and 452 (26.41%) sorghum genes from these regions were tagged by corresponding sequence matches of WCA1H only but did not exhibit any detectable rice/sorghum orthologous counterpart. Thus, we were able to tentatively allocate

1,858 nonredundant gene loci with associated putative rice/sorghum orthologs on barley chromosome 1H. In addition, 129 map-anchored barley loci without corresponding rice/sorghum ortholog have also been integrated into the 1H gene map. This increased the number of oriented and anchored loci to 1,987, which corresponded to between 34% and 43% of the estimated gene complement of chromosome 1H (Supplemental Tables S2 and S4).

The syntenic integration based on information of rice and sorghum provided specifically added value for regions with limited genetic resolution of barley chromosome 1H, i.e. centromeric and subcentromeric regions. Here, sequence identity to collinearly organized homologs (orthologs) of rice and sorghum provided a hypothetical linear order for such barley markers/genes for which linear gene/marker order could not be resolved genetically. Furthermore, the collinear intervals in rice and sorghum that could be framed by cosegregating markers of the barley 1H centromere were carrying as many as 373 genes that

Table III. Repeat content and composition in WCAall and WCA1H datasets

Sequences from the WCAall as well as the WCA1H collection were analyzed for their repeat content. Similar frequencies of each category were observed in the two collections.

Type of Repetitive Element	WCAall	WCA1H
	% of genome	% of genome
Class I: retroelement	67.61	71.10
LTR retrotransposon	66.99	70.41
Ty1/copia	13.41	14.44
Ty3/gypsy	36.44	38.56
Unclassified LTR	17.14	17.41
Non-LTR retrotransposon	0.61	0.68
LINE	0.60	0.67
SINE	0.01	0.01
Unclassified retroelement	0.01	0.01
Class II: DNA transposon	6.44	6.00
DNA transposon superfamily	6.06	5.62
CACTA superfamily	5.59	5.19
hAT superfamily	0.05	0.06
Mutator superfamily	0.24	0.22
Tc1/Mariner superfamily	0.08	0.03
PIF/Harbinger	0.10	0.12
Unclassified	0.01	0.01
DNA transposon derivative	0.24	0.26
MITE	0.24	0.26
Helitron	0.09	0.06
Unclassified DNA transposon	0.05	0.06
High copy number gene	0.13	0.04
RNA gene	0.13	0.04
Total	74.54	77.49

were tagged by WCA1H reads. Given that only between 34% to 43% genes are potentially syntenic between barley, rice, and sorghum in this region (see above) it can be postulated that between 850 to 1,100 genes, roughly 20% of all genes of barley 1H, may be located in centromeric and subcentromeric regions exhibiting very low recombination frequency and thus represent genes with limited accessibility based on genetic mapping approaches.

DISCUSSION

A complete genome sequence is a fundamental resource to answer a wide range of basic and applied scientific questions. However, for the Triticeae tribe comprising some of the most important crop species (i.e. wheat, barley), large-scale genomic sequence information is essentially lacking. Whole genome sequencing of barley and wheat is complicated by the huge genome size (1 C to approximately 5.1 Gbp in barley; Doležel et al., 1998; and 1 C to approximately 17 Gbp in wheat; Bennett and Smith, 1976) and the inherent genome complexity caused by a content of 80% to 90% repetitive elements (Smith and Flavell, 1975; Paux et al., 2006). In this study, we combined chromosome sorting and NGS to gain insight at unprecedented density into the gene content of an entire

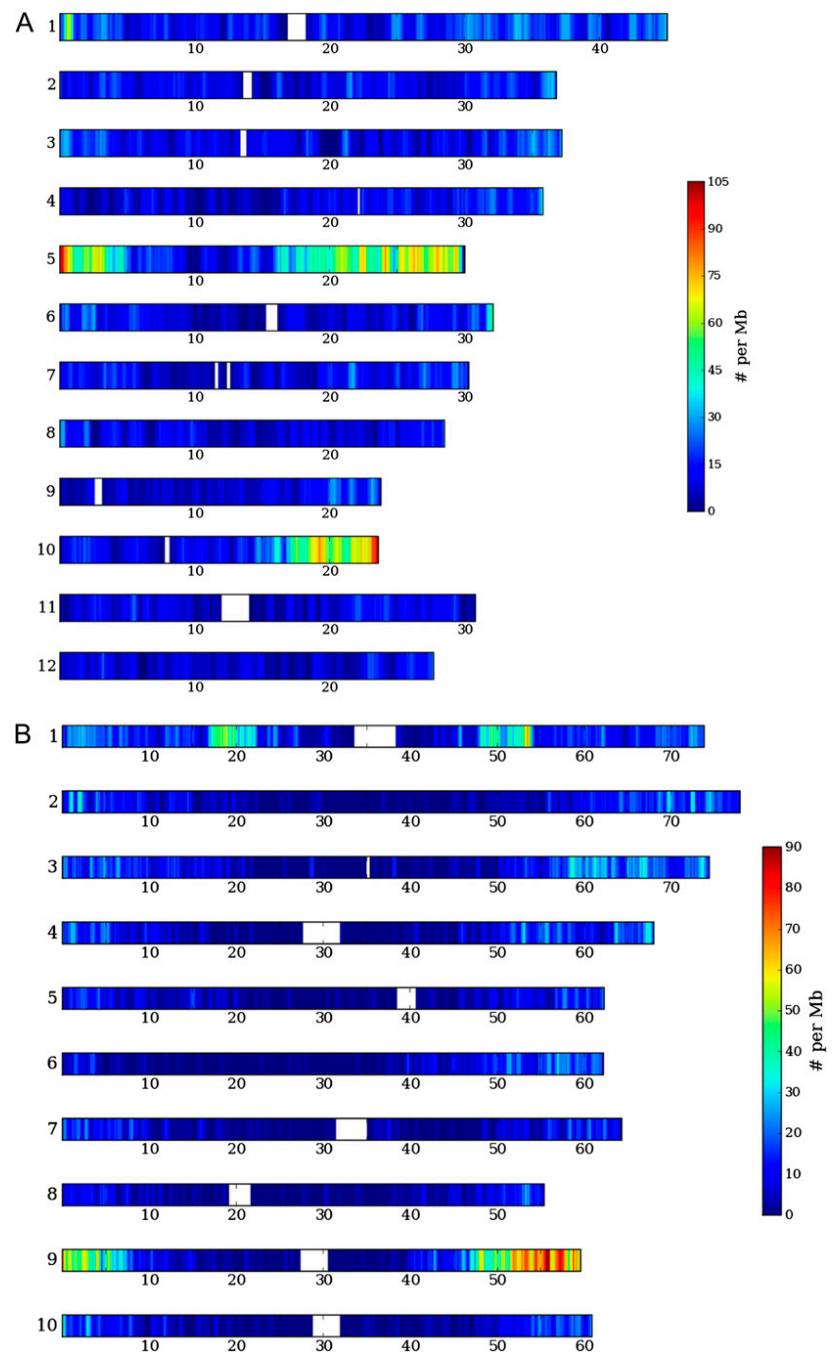
Triticeae chromosome. Integration with high-resolution synteny data from grass model genome sequences of rice and sorghum allowed us to propose a virtually ordered gene inventory of 1,987 anchored genes (39% of sequence-tagged genes) of barley chromosome 1H.

Almost 90% of all genes of chromosome 1H were sequence tagged at only 1.3-fold 454 shotgun sequence coverage. Based on the number of genes detected by 454 sequence reads in the genome reference datasets of rice and sorghum and EST datasets of wheat and barley and a 95% probability of chromosome 1H origin, this translated into a gene content of roughly 5,400 genes for chromosome 1H. Overall 45,000 genes for the entire barley genome can be estimated. This number is very close to a previous estimate based on assembly of 444,652 barley ESTs (28,001 EST contigs + 22,937 EST singles, <http://www.harvest-web.org>; Close et al., 2008) but it slightly exceeds the annotated gene content of rice (37,544 predicted genes; International Rice Genome Sequencing Project, 2005) and sorghum (34,496 gene models; Paterson et al., 2009). Additional indirect confirmation of our gene content estimate came from end sequencing of approximately 11,000 chromosome-specific BAC clones that suggested a content of 6,000 genes for wheat chromosome 3B (Paux et al., 2006). This wheat chromosome is homologous to barley chromosome 3H (size 755 Mb; Suchankova et al., 2006). Assuming a comparable gene density for both barley chromosomes 1H and 3H, the estimated gene content scales to a content of 6,500 genes for barley chromosome 3H, a similar range of magnitude as estimated for wheat chromosome 3B.

Grass genomes share a significant level of synteny (Moore et al., 1995). Colinearity of Triticeae group 1 chromosomes was recently confirmed to distal regions of both arms of rice chromosome 5 and the distal part of rice chromosome 10 long arm on the basis of several hundred gene-derived markers in barley (Stein et al., 2007) and wheat (Qi et al., 2004), respectively. Here, our study takes this analysis to the level of a complete chromosome view: About 36.2% of all genes detected for chromosome 1H matched to rice and/or sorghum genes located in colinear regions and thus confirmed previously detected synteny. More importantly, the sequence coverage, the high degree of chromosome purity, and corresponding syntenic coverage enabled to imply the extent of syntenic regions with a per gene resolution. No further regions with conserved gene content to the rice and sorghum genomes were observed.

The integration of low-pass shotgun sequencing information of barley chromosome 1H with the colinear gene order of 1,858 nonredundant orthologous rice and sorghum genes allowed us to propose a virtual sequence-based gene order map of an entire Triticeae chromosome. It is noteworthy that syntenic integration also allowed the ordering of genes in regions with limited genetic resolution such as subcentromeric and centromeric regions. Our results indicated that roughly one-fifth of the genes of barley chromosome

Figure 3. WCA1H sequence reads mapped on the genomes of rice and sorghum. The heatmap is depicting the location of detected rice (A) and sorghum (B) homologous (syntenic) segments. WCA1H sequence reads were anchored on rice and sorghum using BLASTX and the best detectable match. Individual chromosomes were numbered and the size intervals in megabases were given. Regions with conserved gene content to barley chromosome 1H (implied syntenic regions) were obvious and encompassed rice chromosomes 5 and 10 as well as a small region on chromosome 1. For sorghum, similar regions were observed for chromosomes 1 and 9.



1H are possibly located in this region with low recombination frequency. In addition to the currently available sequences of rice and sorghum, genome sequences will soon become available for maize (*Zea mays*; Pennisi, 2008) and *Brachypodium* (<http://www.brachypodium.org/>), of which the latter is evolutionarily considerably closer to barley (Bolot et al., 2009). Such additional information will allow to further refine gene maps derived from low-pass sequencing of flow-sorted chromosomes. Nevertheless, this approach will also meet limitations: Due to translocation of genes in comparison to the synteny scaffolds, an

estimated 50% of the detected barley genes cannot be anchored and local rearrangements as well as local duplications like tandemly duplicated genes cannot be resolved. Thus, the presented approach can be seen as a powerful approximation and as a complementary approach to other genetic and physical map-based attempts to develop a complete reference genome sequence of barley and Triticeae in general.

Flow cytometric sorting provides a powerful means to reduce genome complexity since it allows isolation of individual chromosomes (Doležel et al., 2007). In our study we focused on barley chromosome 1H

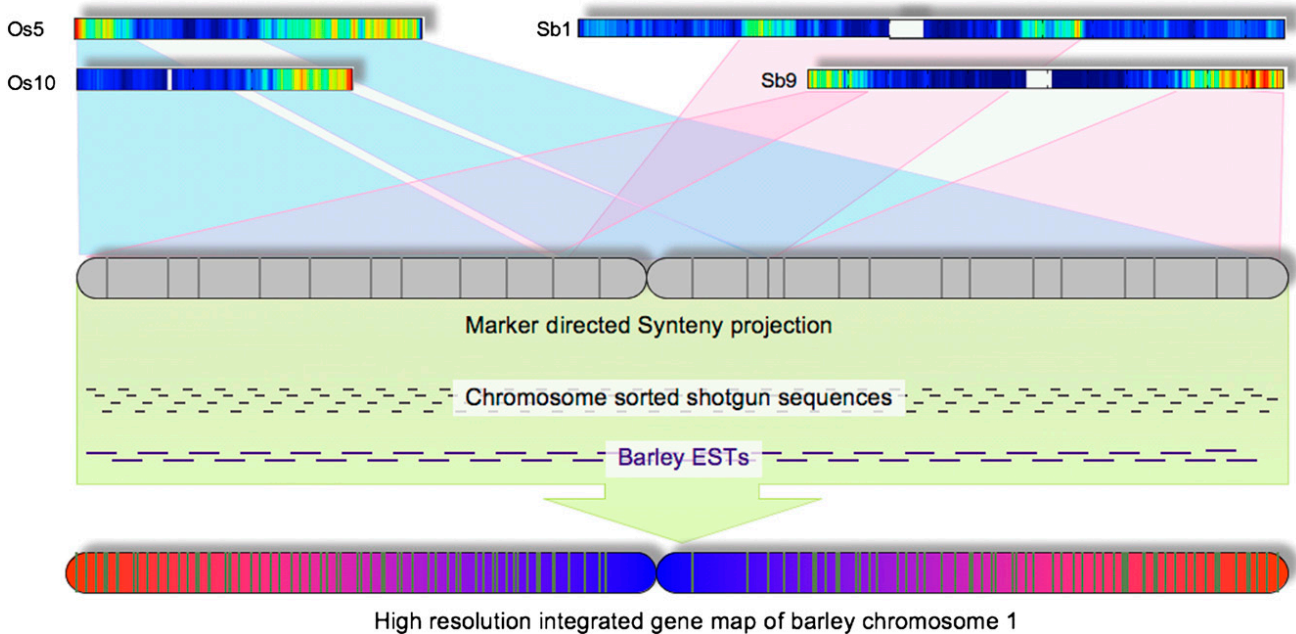


Figure 4. Schematic representation of marker and synteny guided assembly of an integrated virtual gene map for barley chromosome 1H. Genetically anchored barley markers have been integrated with rice and sorghum genes located in syntenic regions to give an enriched tentative ancestral gene scaffold. WCA1H sequence reads as well as barley EST sequences have been associated with this chromosome matrix and give rise to an ordered integrated gene map of barley chromosome 1H.

(approximately 622 Mb), which represents about 12% of the barley genome and that can be directly sorted from the remaining six chromosomes (Suchankova et al., 2006). The remaining barley chromosomes 2H to 7H can be sorted separately from wheat-barley ditelosomic addition lines (Suchankova et al., 2006). Such chromosome arms represent between 6% and 9% of the barley genome (301–459 Mbp) and would enable to survey the whole barley genome by NGS low-pass shotgun sequencing at further reduced complexity.

In this study, low-pass shotgun sequencing of flow-sorted chromosomes proved to be efficient to sequence tag the gene content of a whole barley chromosome. Instead of direct sequencing of chromosomal DNA, MDA (Dean et al., 2002) was used to generate microgram quantities of DNA from batches of 10,000 sorted 1H chromosomes. MDA has proven to be useful for highly accurate and representative amplification of human, fungal, and microbial templates (Silander and Saarela, 2008) as well as for flow-sorted barley chromosomes (Simkova et al., 2008). The potential value of this source of DNA for de novo shotgun sequencing and for genome sequence assembly in the Triticeae, however, remains to be determined.

De novo shotgun sequencing has been previously applied to moderately complex plant genomes that exceed the size of individual barley chromosomes and harbor tracks of highly repetitive sequences in the range of several megabases. So far such attempts either relied on Sanger sequencing only or used Sanger and NGS technology in mixed assemblies (Jaillon et al., 2007; Velasco et al., 2007; Paterson et al., 2009). In all

cases, however, paired-end sequencing of differently but specifically sized DNA fractions (i.e. genomic plasmid, cosmid, or BAC libraries) was applied to obtain sufficiently sized sequence scaffolds. Since MDA DNA contains a low-amplification bias (Dean et al., 2002; Hosono et al., 2003; Rook et al., 2004) the method might contribute to upcoming strategies for whole chromosome and genome shotgun sequencing and assembly in Triticeae.

CONCLUSION

Low-pass shotgun sequencing of flow-sorted barley chromosome 1H boosted the amount of 1H anchored genes by 6-fold compared to existing map resources. With the integration of syntenic information from other grass genomes unprecedented resolution was achieved. This data will significantly impact cereal genomics: Anchored as well as the unanchored genes determined in this study can be correlated with BAC clone libraries and thus anchored to the emerging physical map of the barley genome (Schulte et al., 2009). In prospect of the rapid improvement of sequencing technology (Shendure and Ji, 2008) and upcoming highly advanced genomic resources for the Triticeae (dense marker frameworks, robust physical maps, reduced DNA sample complexity by chromosome sorting, access to syntenic reference grass genome sequences) the cost-effective generation of sequences for individual chromosome arms and finally the complete barley genome is no longer far out of reach.

MATERIALS AND METHODS

Purification and Amplification of Chromosomal DNA

Intact mitotic chromosomes were isolated by flow cytometric sorting and the purity of the obtained chromosome suspension was determined by FISH essentially as described previously (Suchankova et al., 2006). The DNA of sorted chromosomes was purified and amplified by MDA as described by Šimková et al. (2008).

454 Sequencing

DNA amplified from sorted chromosome 1H (WCA1H) and from sorted chromosomes 1H to 7H (WCAall) was used for 454 shotgun sequencing. Five micrograms of MDA DNA was used to prepare the 454 sequencing library using the GS FLX DNA library preparation kit, following the manufacturer's instructions (Roche Diagnostics). Single-stranded 454 sequencing libraries were quantified by a quantitative PCR assay (Meyer et al., 2008) and processed utilizing a GSFLX standard emPCR kit I and standard LR70 sequencing kit (Roche Diagnostics) according to manufacturer's instructions. For WCA1H, six complete GS FLX sequencer runs (70 × 75 picotiter plates) resulted in 3,046,327 reads with a median read length of 258 bp, yielding 799,343,261 bp of raw sequence data (675,561,265 high-quality bases). Two runs with DNA from pooled chromosomes 1H to 7H (WCAall) using half of a 70 × 75 picotiter plate resulted in overall 381,617 reads (median read length = 259 bp), yielding 99,401,554 bp raw sequence data (90,536,939 high-quality bases). Sequencing details were summarized in Table I. All sequence information generated in this study was submitted to the National Center for Biotechnology Information short read archive under accession number SRP001030.

Sequence Analysis

Analysis of Repetitive DNA and Repeat Masking of Sequences

Initially the content of repetitive DNA per sequence read was identified by analysis with RepeatMasker (<http://www.repeatmasker.org>) against the MIPS-REdat Poaceae v8.1 repeat library (contains known grass transposons from the Triticeae Repeat Database, <http://wheat.pw.usda.gov/ITMI/Repeats>, as well as de novo detected LTR retrotransposon sequences from several grass species, e.g. maize [*Zea mays*]: 12,434, sorghum [*Sorghum bicolor*]: 7,500, rice [*Oryza sativa*]: 1,928, *Brachypodium distachyon*: 466, wheat [*Triticum aestivum*]: 356, and barley [*Hordeum vulgare*]: 86 sequences). Subsequently, repetitive regions were masked by vmatch (<http://www.vmatch.de>) at the following parameters: 55% identity cutoff, 30 bp minimal length, seed length 14, exdrop 5, *e* value 0.001.

Sequence-Tagged Genes in the WCA1H Sequence Dataset

To estimate the number of barley genes that have been captured in the WCA1H sequence collection, BLAST (Altschul et al., 1990) comparisons were carried out with the repeat-filtered reads against the rice and sorghum proteins/coding sequences as well as against clustered wheat and barley EST collections (HarvEST, <http://harvest.ucr.edu/>; barley v1.73, assembly 35, wheat v1.16; Rice RAP-DB genome build 4, <http://rapdb.dna.affrc.go.jp>; sorghum genome annotation v1.4 [<http://genome.jgi-psf.org/Sorbi1/Sorbi1.download.ftp.html>]; Paterson et al., 2009). The number of tagged genes and the number of gene matching reads were counted after filtering according to the following criteria: (1) the best hit display with a similarity greater than an adjusted species-specific similarity characteristic (see below for definition) and (2) an alignment length ≥ 30 amino acids (BLASTN 50 bp). A species-adapted similarity cutoff value was calibrated before by performing similarity searches (BLASTX/TBLASTX/BLASTN) of barley EST clusters against rice and sorghum proteins and against wheat ESTs/tentative consensi (similarity cutoff: sorghum 75%, rice 80%, wheat 85%; see Supplemental Fig. S1, A and B).

Identification of Genetic Markers in the WCA1H and WCAall Datasets

The repeat-masked sequence collections from WCA1H and WCAall were compared (BLASTN) against 2,785 nonredundant (of total 2,943) EST-based

markers (<http://harvest.ucr.edu>) under optimized parameters (-r 1 -q -1 -W 9 -G 1 -E 2: -r reward for a nucleotide match, default = 1; -q penalty for a nucleotide mismatch, default = -3; -G cost to open a gap, default = -1; -E cost to extend a gap, default = -1; -W word size, default). Only BLAST matches exceeding a similarity threshold of 98% and an alignment length ≥ 50 bp were further analyzed.

Comparative Genomics to Rice and Sorghum and Syntenic Integration

The WCA1H dataset was compared (BLASTX) to the reference genomes of rice and sorghum at a filter criterion of ≥ 30 amino acid similarity. Matched rice and sorghum genes were plotted along their position on the respective chromosomes and the average syntenic content (number of WCA1H matched genes per window size of 10 genes in rice and sorghum, respectively) was computed and visualized in heatmaps.

All rice and sorghum genes contained in syntenic regions in barley that could be delimited by a scaffold of 332 barley chromosome 1H-allocated EST-based markers and that exhibited a match to individual WCA1H 454 sequence reads were selected and integrated, producing a syntenic scaffold. First, putatively orthologous rice and sorghum genes were determined in this set of genes by reciprocal BLASTP searches considering only best matches. Subsequently, genes present either only in rice or sorghum but exhibited matches to WCA1H 454 reads were sorted in between.

All sequence information generated in this study was submitted to the NCBI GenBank short read archive under accession number SRP001030.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Sequence comparisons of barley ESTs against wheat, rice, and sorghum genes.

Supplemental Table S1. Sequence similarities in coding regions between the genomes of rice and sorghum and EST resources from wheat and barley.

Supplemental Table S2. Reconstruction of barley chromosome 1H by using syntenic relationships.

Supplemental Table S3. Comparison of barley chromosome 1H enriched sequences (WCA1H) with chromosomes of rice and sorghum.

Supplemental Table S4. Virtual gene order list of barley chromosome 1H based on syntenic integration.

ACKNOWLEDGMENTS

We are grateful to Dr. Z. Stehno (Crop Research Institute, Prague, Czech Republic) for providing seeds of barley cv Morex and we kindly acknowledge the excellent technical assistance of D. Werler, I. Heinze, and C. Luge as well as D. Riano-Pachon from www.gabipd.org for support in sequence data submission.

Received June 7, 2009; accepted August 13, 2009; published August 19, 2009.

LITERATURE CITED

- Altschul SE, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410
- Bennett MD, Smith JB (1976) Nuclear DNA amounts in angiosperms. *Philos Trans R Soc Lond B Biol Sci* 274: 227–274
- Bennetzen JL, Freeling M (1993) Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genet* 9: 259–261
- Bolot S, Abrouk M, Masood-Quraishi U, Stein N, Messing J, Feuillet C, Salse J (2009) The 'inner circle' of the cereal genomes. *Curr Opin Plant Biol* 12: 119–125
- Close TJ, Wanamaker S, Roose ML, Lyon M (2008) HarvEST. *In* D Edwards, ed, *Plant Bioinformatics*, Vol 406. Humana Press, New York, pp 161–177

- Dean FB, Hosono S, Fang L, Wu X, Faruqi AF, Bray-Ward P, Sun Z, Zong Q, Du Y, Du J, et al (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci USA* **99**: 5261–5266
- Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R (1998) Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann Bot (Lond) (Suppl A)* **82**: 17–26
- Doležel J, Kubaláková M, Paux E, Bartoš J, Feuillet C (2007) Chromosome-based genomics in the cereals. *Chromosome Res* **15**: 51–66
- Draper J, Mur LA, Jenkins G, Ghosh-Biswas GC, Bablak P, Hasterok R, Routledge AP (2001) *Brachypodium distachyon*: a new model system for functional genomics in grasses. *Plant Physiol* **127**: 1539–1555
- Gaut BS (2002) Evolutionary dynamics of grass genomes. *New Phytol* **154**: 15–28
- Hosono S, Faruqi AF, Dean FB, Du Y, Sun Z, Wu X, Du J, Kingsmore SF, Egholm M, Lasken RS (2003) Unbiased whole-genome amplification directly from clinical samples. *Genome Res* **13**: 954–964
- International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature* **436**: 793–800
- Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**: 463–467
- Komatsuda T, Pourkheirandish M, He C, Azhaguvel P, Kanamori H, Perovic D, Stein N, Graner A, Wicker T, Tagiri A, et al (2007) Six-rowed barley originated from a mutation in a homeodomain-leucine zipper I-class homeobox gene. *Proc Natl Acad Sci USA* **104**: 1424–1429
- Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* **2**: 231–239
- Linde-Laursen IB (1996) Recommendations for the designation of the barley chromosomes and their arms. *Barley Genet Newsl* **26**: 1–3
- Marthe F, Künzel G (1994) Localization of translocation breakpoints in somatic metaphase chromosomes of barley. *Theor Appl Genet* **89**: 240–248
- Meyer M, Briggs AW, Maricic T, Hober B, Hoffner B, Krause J, Weihmann A, Paabo S, Hofreiter M (2008) From micrograms to picograms: quantitative PCR reduces the material demands of high-throughput sequencing. *Nucleic Acids Res* **36**: e5
- Moore G, Devos KM, Wang Z, Gale MD (1995) Cereal genome evolution: grasses, line up and form a circle. *Curr Biol* **5**: 737–739
- Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556
- Paux E, Roger D, Badaeva E, Gay G, Bernard M, Sourdille P, Feuillet C (2006) Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J* **48**: 463–474
- Pennisi E (2008) Plant sciences: corn genomics pops wide open. *Science* **319**: 1333
- Qi LL, Echalié B, Chao S, Lazo GR, Butler GE, Anderson OD, Akhunov ED, Dvorák J, Linkiewicz AM, Ratnasiri A, et al (2004) A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**: 701–712
- Rook MS, Delach SM, Deyneko G, Worlock A, Wolfe JL (2004) Whole genome amplification of DNA from laser capture-microdissected tissue for high-throughput single nucleotide polymorphism and short tandem repeat genotyping. *Am J Pathol* **164**: 23–33
- Sasaki T, Sederoff RR (2003) The rice genome and comparative genomics of higher plants. *Curr Opin Plant Biol* **6**: 97–100
- Schulte D, Close TJ, Graner A, Langridge P, Matsumoto T, Muehlbauer G, Sato K, Schulman AH, Waugh R, Wise RP, et al (2009) The international barley sequencing consortium—at the threshold of efficient access to the barley genome. *Plant Physiol* **149**: 142–147
- Shendure J, Ji H (2008) Next-generation DNA sequencing. *Nat Biotechnol* **26**: 1135–1145
- Silander K, Saarela J (2008) Whole genome amplification with Phi29 DNA polymerase to enable genetic or genomic analysis of samples of low DNA yield. *Methods Mol Biol* **439**: 1–18
- Simkova H, Svensson JT, Condamine P, Hribova E, Suchankova P, Bhat PR, Bartos J, Safar J, Close TJ, Dolezel J (2008) Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* **9**: 294
- Singh RJ, Tsuchiya T (1982) An improved Giemsa n-banding technique for the identification of barley chromosomes. *J Hered* **73**: 227–229
- Smith DB, Flavell RB (1975) Characterisation of the wheat genome by renaturation kinetics. *Chromosoma* **50**: 223–242
- Stein N (2007) Triticeae genomics: advances in sequence analysis of large genome cereal crops. *Chromosome Res* **15**: 21–31
- Stein N, Prasad M, Scholz U, Thiel T, Zhang H, Wolf M, Kota R, Varshney RK, Perovic D, Grosse I, et al (2007) A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor Appl Genet* **114**: 823–839
- Suchankova P, Kubaláková M, Kovarova P, Bartos J, Cihalikova J, Molnar-Lang M, Endo TR, Dolezel J (2006) Dissection of the nuclear genome of barley by chromosome flow sorting. *Theor Appl Genet* **113**: 651–659
- Velasco R, Zharkikh A, Troglio M, Cartwright DA, Cestaro A, Pruss D, Pindo M, Fitzgerald LM, Vezzulli S, Reid J, et al (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. *PLoS One* **2**: e1326
- Wicker T, Narechania A, Sabot E, Stein J, Vu GT, Graner A, Ware D, Stein N (2008) Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9**: 518
- Wicker T, Schlagenhauf E, Graner A, Close TJ, Keller B, Stein N (2006) 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**: 275
- Wicker T, Taudien S, Houben A, Keller B, Graner A, Platzer M, Stein N (2009) 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* (in press)
- Yan L, Loukoianov A, Blechl A, Tranquilli G, Ramakrishna W, SanMiguel P, Bennetzen JL, Echenique V, Dubcovsky J (2004) The wheat *VRN2* gene is a flowering repressor down-regulated by vernalization. *Science* **303**: 1640–1644

LARGE-SCALE BIOLOGY ARTICLE

Unlocking the Barley Genome by Chromosomal and Comparative Genomics

Klaus F.X. Mayer,^{a,1} Mihaela Martis,^a Pete E. Hedley,^b Hana Šimková,^c Hui Liu,^b Jenny A. Morris,^b Burkhard Steuernagel,^d Stefan Taudien,^e Stephan Roessner,^a Heidrun Gundlach,^a Marie Kubaláková,^c Pavla Suchánková,^c Florent Murat,^f Marius Felder,^e Thomas Nussbaumer,^a Andreas Graner,^d Jerome Salse,^f Takashi Endo,^g Hiroaki Sakai,^h Tsuyoshi Tanaka,^h Takeshi Itoh,^h Kazuhiro Sato,ⁱ Matthias Platzer,^e Takashi Matsumoto,^h Uwe Scholz,^d Jaroslav Doležel,^c Robbie Waugh,^{b,1} and Nils Stein^{d,1,2}

^a Munich Information Center for Protein Sequences/Institute of Bioinformatics and Systems Biology, Institute for Bioinformatics and Systems Biology, Helmholtz Center Munich, 85764 Neuherberg, Germany

^b Scottish Crop Research Institute, Invergowrie, Dundee, Scotland DD25DA, United Kingdom

^c Centre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, 77200 Olomouc, Czech Republic

^d Leibniz Institute of Plant Genetics and Crop Plant Research, 06466 Gatersleben, Germany

^e Leibniz Institute for Age Research-Fritz Lipmann Institute, 07745 Jena, Germany

^f Institut National de la Recherche Agronomique Clermont-Ferrand, Unité Mixte de Recherche, Institut National de la Recherche Agronomique, Université Blaise Pascal 1095, Amélioration et Santé des Plantes, Domaine de Crouelle, Clermont-Ferrand 63100, France

^g Kyoto University, Laboratory of Plant Genetics, Kyoto 606-8502, Japan

^h National Institute of Agrobiological Sciences, Tsukuba, Ibaraki 305-8602, Japan

ⁱ Okayama University, Institute of Plant Science and Resources, Kurashiki 710-0046, Japan

We used a novel approach that incorporated chromosome sorting, next-generation sequencing, array hybridization, and systematic exploitation of conserved synteny with model grasses to assign ~86% of the estimated ~32,000 barley (*Hordeum vulgare*) genes to individual chromosome arms. Using a series of bioinformatically constructed genome zippers that integrate gene indices of rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), and *Brachypodium distachyon* in a conserved synteny model, we were able to assemble 21,766 barley genes in a putative linear order. We show that the barley (H) genome displays a mosaic of structural similarity to hexaploid bread wheat (*Triticum aestivum*) A, B, and D subgenomes and that orthologous genes in different grasses exhibit signatures of positive selection in different lineages. We present an ordered, information-rich scaffold of the barley genome that provides a valuable and robust framework for the development of novel strategies in cereal breeding.

INTRODUCTION

Access to a genome sequence is now considered pivotal for unraveling key questions in crop plant biology and interrogating the molecular mechanisms that underpin trait formation. A genome sequence is central to the development of true genomics-informed breeding strategies and for unlocking the full potential of natural genetic variation for future crop improvement. Unfor-


tunately for several key crops, deciphering a complete genome sequence to date has been precluded by the size and/or complexity of their genomes. Given the combined challenges of food security and climate change, it is vital that this situation is resolved and resources are developed that, even if not meeting an optimal gold standard, in the interim provide a high value and high utility surrogate.


Despite their importance in global agriculture, the Triticeae species wheat (*Triticum aestivum*; $2n=6x=42$) and barley (*Hordeum vulgare*; $2n=2x=14$), ranked 1 and 5 in world food production (FAOSTAT, 2007; <http://faostat.fao.org/>), are two such crops where genome size and complexity (17 Gbp for wheat [Bennett and Smith, 1976] and 5.1 Gbp for barley [Doležel et al., 1998]) so far preclude the development of such a gold standard reference genome sequence. Genomic data both from sequenced BAC clones and the application of next-generation sequencing (NGS) methodologies are available at a limited scale (Steuernagel et al.,

¹ These authors contributed equally to this work.

² Address correspondence to stein@ipk-gatersleben.de.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Nils Stein (stein@ipk-gatersleben.de).

 Online version contains Web-only data.

 Open Access articles can be viewed online without a subscription. www.plantcell.org/cgi/doi/10.1105/tpc.110.082537

2009; Wicker et al., 2009; <http://www.cerealsdb.uk.net/>) but lack the context required for broad and general utility. Given a close evolutionary relationship (divergence 13 million years ago [MYA]; Gaut, 2002) that has resulted in extensive conservation of synteny (Moore et al., 1995; Devos, 2005), it is generally accepted that elucidating a genome sequence for barley, a genetically tractable diploid inbreeder, would serve both its own genetics and breeding communities well while providing a faithful proxy for the genomically taxing 17 Gbp hexaploid bread wheat genome. This proposition is supported by agronomic traits such as flowering time and vernalization response being shared with wheat and the causal genes located at conserved genomic regions (Fu et al., 2005; Turner et al., 2005; Yan et al., 2006; Beales et al., 2007). Even race-specific disease resistance, a paradigm for species-specific genetic control in plants, shares conserved genetic elements in barley and wheat. Recently, a functional allele of the barley gene *Mla*, which confers resistance to the powdery mildew fungus (Zhou et al., 2001), was isolated from *Triticum monococcum* (Jordan et al., 2010). Indeed, an increasing body of information supports the notion of treating the Triticeae as a single genetic system.

Barley is itself an important crop. In addition to being the raw material for the brewing and distilling industry, barley is an important component of animal feed, can contribute health benefits in the human diet, and is agroecologically important, being planted worldwide on >57 million hectares (FAOSTAT, 2010; <http://www.fao.org/faostat>), often as an integral component of crop rotation management. Historically, it also has been an important model for classical genetics where its diploid genome has facilitated genetic analysis, a position that extended into the genomics era where early EST sequences provided resources for microarray design that in turn established routine functional genomics (Close et al., 2004; Druka et al., 2006). Subsequently, the same sequences were exploited to generate high-density gene maps using innovative marker technology (Stein et al., 2007; Potokina et al., 2008; Close et al., 2009; Sato et al., 2009a), and these opened the way for in-depth comparative analyses with other grass genomes (Bolot et al., 2009; Thiel et al., 2009; Abrouk et al., 2010; Murat et al., 2010). More recently, detailed information about barley genome composition has been accumulated using NGS technologies (Wicker et al., 2006, 2008, 2009). Despite the significance of each of these advances, the difficulties associated with fully unraveling the complex and repeat-rich 5.1-Gbp barley genome remain a significant challenge.

Recently, we demonstrated the potential of a cost-efficient and integrated cytogenetics, molecular genetics, and bioinformatics approach for generating a specific gene index for an entire barley chromosome. From a Roche 454 data set of 1.3-fold coverage generated from flow-sorted barley chromosome 1H, sequence signatures of >5000 genes were extracted and integrated with data from the rice (*Oryza sativa*) and sorghum (*Sorghum bicolor*) genomes to deliver a comprehensive virtual linear gene order model (Mayer et al., 2009). Here, we extended this approach by incorporating full-length cDNA (fl-cDNA) and DNA hybridization microarray data and applied it to the whole barley genome. This has allowed us to develop the first blueprint of a diploid Triticeae genome: a genome-wide putative linear

gene index of barley embedded in a comparative grass genome organization model. The model is founded in an assembled series of genome zippers, a bioinformatics framework that exploits the extensive conservation of synteny observed between fully sequenced grass genomes.

RESULTS

Gene Content of Barley

We purified separately an entire barley chromosome (1H) and 12 chromosome arms (2HS to 7HL) by flow cytometry, amplified the DNA by multiple displacement amplification (MDA), and then shotgun sequenced the resulting preparations to 1.04- to 2.00-fold coverage using Roche 454 technology (Table 1; see Supplemental Table 1 online). At this depth of sequencing, base pair coverage for the individual samples was estimated to range between 64.7 and 86.5% according to Lander-Waterman genome assembly statistics (Lander and Waterman, 1988). We tested this estimate by comparing the individual sequence collections against a genetic map comprised of 2785 nonredundant gene-based single nucleotide polymorphism markers (Close et al., 2009). The observed gene (marker) discovery rate (i.e., the sensitivity) from individual chromosome arms ranged from 81.0 to 98.0% (average sensitivity of 85.9%; see Supplemental Data Set 1 online) exceeding the estimated values.

We then assessed the purity of the chromosome/chromosome arm fractions by counting the proportion of false positive and true negative matches in the data set (i.e., the specificity). Specificities ranged from 88 to 98% (average 96.8%; see Supplemental Data Set 1 online). Applying a confusion matrix, the probability for correct classification reached between 0.89 and 0.97 (average 0.96) for individual chromosome arms (see Supplemental Table 2 online). These findings are consistent with a purity of enrichment estimated by fluorescent in situ hybridization analysis of the individual sorted chromosomal fractions (see Supplemental Table 3 online). Overall the data indicated >95% confidence that genes detected in a chromosome arm sequence data set originated from the assigned source.

To both validate and extend the 454 sequencing-based observations, we generated a complementary chromosome arm gene content data set by hybridizing individual preparations (in three replications) to barley long-oligonucleotide microarrays. In total, we were able to assign 16,804 genes on the array to individual chromosome arms at high confidence (see Supplemental Figure 1 online). Using the previously defined criteria, the genes assigned by array hybridization revealed an average specificity of 99%.

Given the high purity of the flow sorted chromosome samples, we attempted to determine a minimum set of genes for the barley genome. Both 454 sequence and array hybridization-based data sets were compared against complete model grass genomes using BLASTX (similarity $\geq 75\%$ and ≥ 30 amino acids). From the 454 data, 17,290, 18,340, and 19,289 genes were detected from rice, sorghum, and *Brachypodium distachyon*, respectively, resulting in a cumulative set of 21,240 nonredundant homologous genes (Table 2). Sequence comparison of the 16,804 array-based

Table 1. Sequence and Coverage Statistics of Individual Barley Chromosomes and Chromosome Arms

Chromosome/ Chromosome Arm	Size (Mbp)	Sequences (Mbp)	Sequences of High Quality (Mbp)	Reached Coverage (X-Fold)	Reached Coverage of High-Quality Sequences (X-Fold)	Expected Lander Waterman	Expected Lander Waterman of High-Quality Sequences	Observed Marker Detection Rate (Sensitivity) of High-Quality Sequences
1H Morex	622	798	675	1.28	1.09	72.00%	66.38%	95.18
1H Betzes	622	813	569	1.31	0.91	73.01%	59.74%	88.55
1H (MoBe)	622	1,611	1,244	2.60	2.00	92.57%	86.46%	98.19
2HS	362	528	377	1.46	1.04	76.78%	64.65%	82.35
2HL	428	924	670	2.16	1.57	88.47%	79.20%	86.24
3HS	336	657	470	1.96	1.40	85.91%	75.34%	80.58
3HL	419	1,155	744	2.76	1.78	93.67%	83.14%	85.95
4HS	336	653	452	1.94	1.35	85.63%	74.08%	80.55
4HL	393	911	605	2.32	1.54	90.17%	78.56%	83.01
5HS	301	760	546	2.52	1.81	91.95%	83.63%	90.29
5HL	459	949	651	2.07	1.42	87.38%	75.83%	83.03
6HS	332	830	570	2.50	1.72	91.79%	82.09%	86.29
6HL	357	981	587	2.75	1.64	93.61%	80.60%	86.38
7HS	382	640	505	1.67	1.32	81.17%	73.29%	80.97
7HL	373	636	468	1.70	1.25	81.73%	71.35%	84.89
	(Σ) 5,100	(Σ) 11,235	(Σ) 7,889	($\bar{\Sigma}$) 2.20	($\bar{\Sigma}$) 1.55	($\bar{\Sigma}$) 88.91%	($\bar{\Sigma}$) 78.77%	($\bar{\Sigma}$) 86.16

Basic statistics for chromosome (arm)-based shotgun sequencing of the barley genome. The table lists individual chromosome (arm) sizes, sequence data generated, coverage reached, the theoretical coverage as defined by the Lander Waterman equation, and the marker detection rate for the individual chromosome (arms). The accession used for sequencing was barley cultivar Betzes. For chromosome 1H, data previously generated in the barley cultivar Morex (Mayer et al., 2009) were combined with data generated in the cv Betzes. Statistics are given for the individual cultivars as well as the combined data set. Summary values given are from the combined Morex/Betzes data rather than the individual data sets.

unigenes assigned to barley chromosome arms identified an overlapping set of 11,708 genes that were also detected in the 454 sequence data. In total, 10,865 (93%) provided the same chromosomal assignment, consistent with chromosome purity estimates. Of these, 5096 genes were exclusively detected by microarray hybridization leading to an additional 3357, 3438, and 3908 homologous genes identified in rice, sorghum, and *Brachypodium*, respectively (totaling 4046 nonredundant genes) (Table 2). Thus, a cumulative set of 25,286 genes was detected by comparing 454 sequence and array-based data against all three model genomes (Table 2).

To determine how many barley genes can be detected in the three model genomes by stringent homology searches, we used a set of 23,588 nonredundant barley fl-cDNAs. These can be considered as an unbiased reference that represent randomly selected complete coding sequence of genes. In total, 5384 fl-cDNA's remained without a corresponding match (similarity \geq 75%, length \geq 30 amino acids). Thus, some 23% of all barley genes lack sufficient sequence similarity to any gene of the three model grass genomes (Table 2). This is consistent with the value found for the hybridization-based results indicating that the array-based unigene set is a representative collection. Taking the 25,286 nonredundant barley genes detected from 454 and array-based data together with 5384 fl-cDNA that do not match homologs in the three model genomes gives an overall set of 30,670 sequence-supported barley genes.

Based on the experimental sensitivity of 86% for the 454 sequence data, the maximum cumulative overlap of nonredundant

homologous genes between barley and the three model genomes would increase from 21,240 to 24,698 genes (Table 2). Since only 77% of the barley genes have a homolog in any of the three model genomes of rice, *Brachypodium*, or sorghum at the stringency applied, an overall content of \sim 32,000 ($24,698/77 \times 100$) genes can be postulated for the entire barley genome (Table 2). This is in the range of the gene counts provided for the annotated *Brachypodium*, rice, and sorghum genomes (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; The International Brachypodium Initiative, 2010). In summary, we estimate that as many as 96% (30,670/32,000) of the barley gene repertoire is represented by either 454 sequence data, array-based unigenes, or fl-cDNAs used in this study.

A First Draft of the Linear Gene Order in the Barley Genome

To establish a hypothetical order for the genes assigned to chromosome arms, we constructed a multilayered scaffold based on conserved synteny for all barley chromosomes (see Supplemental Figure 2 online). We first identified syntenic regions for each chromosome arm in each of the three model grass genomes by sequence comparison of (repeat-masked) 454 sequences and hybridization probes. Figures 1 and 2 show the comparisons with *Brachypodium* and rice, respectively, and the sorghum comparison is presented in Supplemental Figure 3 online. The respective conserved syntenic regions were selected, and only genes that exhibited a corresponding match from barley 454 sequences and/or hybridization probes were

Table 2. Estimated Gene Content of Barley

Data Sets	Nonredundant Genes			Nonredundant Genes (Cumulative)
	<i>Brachypodium</i>	Rice	<i>Sorghum</i>	
Chr. arm 454 data	19,289	17,290	18,340	21,240
Chr. arm-specific array probes (16,804)	12,382 (74%)	10,617 (63%)	10,915 (65%)	12,755 (76%)
Chr. arm-specific array probes not overlapping with 454 data set (5,196)	3,908 (75%)	3,357 (65%)	3,438 (66%)	4,046 (78%)
Genes detected from 454 data and array hybridization	23,197	20,647	21,778	25,286
Nonredundant fl-cDNA (23,588)	17,622 (75%)	15,340 (65%)	15,419 (65%)	18,204 (77%)
Barley genes detected from 454, array hybridization, and fl-cDNA data	29,163	28,895	29,947	30,670
Estimated number of homologs considering complete genome 454 data	22,429 (85%)	20,104 (71%)	21,325 (77%)	24,698
Number of matching nonredundant fl-cDNA against reference genomes (out of 23,588)	17,622 (75%)	15,340 (65%)	15,419 (65%)	18,204 (77%)
Estimated total (24,698/77 × 100)				32,075

BLASTX comparisons against the reference genomes of *Brachypodium*, rice, and sorghum were undertaken using a stringent filter criterion of $\geq 75\%$ sequence similarity spanning ≥ 30 amino acids. Sequence-tagged genes of barley deduced from similarity comparisons of Roche 454, array-based, and fl-cDNA data sets against reference genomes.

used for integration into the barley scaffold. The mapped and ordered barley gene-based marker map comprising 2785 markers (Close et al., 2009) formed the integration scaffold for the detected orthologous genes and formed a genome-wide framework of sequence-based homology bridges upon which we interlaced all of the intervening genes present in the model genome sequences. Finally, we compiled (i.e., zipped up) the complementary sets of information to form a combined and ordered gene content model for seven barley pseudochromosomes. We call these genome zippers (see Supplemental Data Sets 2 to 8 online). They contain all of the genes in each of the three model species organized on a barley genetic framework associated with the corresponding barley genomic sequence tags, barley ESTs, and barley full-length cDNAs.

By this procedure, between 2261 and 3616 genes were tentatively positioned along each of the individual barley chromosomes, representing a cumulative set of 21,766 genes across the entire barley genome (Table 3, Figures 1 and 2; see Supplemental Figure 3 and Supplemental Data Sets 2 to 8 online). An additional set of 5815 genes could not be integrated into the genome zippers based on conserved synteny models but were associated with individual chromosomes/chromosome arms. Overall, we were able to tentatively position 27,581 barley genes, or 86% of the estimated 32,000 gene repertoire of the barley genome, into chromosomal regions.

Positioning of Barley Centromeres

The genetic centromere of barley chromosomes is characterized by large clusters of genes/markers whose order cannot be genetically resolved due to insufficient recombination in relatively small mapping populations ($n = 100$ to 200). The analysis of DNA samples from individual arms of barley chromosomes 2H to 7H enabled us to deduce the transition from proximal (short) to distal (long) chromosome arms (i.e., the centromere position; see Supplemental Data Sets 2 to 8 online; genome zippers). For

barley 1H, only entire chromosomes could be sorted. However, arm-specific information could be deduced based on available sorted chromosome arm shotgun sequence data of the highly collinear homoeologous chromosome 1A of wheat (T. Wicker, K.F.X. Mayer, and N. Stein, unpublished results). For all chromosomes, a single position (1H = 50 centimorgans [cM], 2H = 59.21 cM, 3H = 55.57cM, 4H = 48.72 cM, 5H = 51.3 cM, 6H = 55.36 cM, and 7H = 78.22 cM) was identified that contained genes allocated by 454 sequence reads to either the short or the long arm DNA data sets. Hence, we defined this to be the genetic position of the respective centromeres and ordered the genes here according to conserved synteny with the genomic models. Among 21,766 genes anchored to the genome zipper, 3125 (14%) genes were allocated to these genetic centromeres. Based on the 454 sequence- and array-based gene assignment to chromosome arms, we could distribute all but nine of these 3125 genes to specific arms of chromosomes 1H to 7H.

A Mosaic of Collinearity Is Observed between Barley and Model Grass Genomes

Shotgun sequencing and array hybridization provided chromosome arm gene content that was translated into tentative linear gene orders using conserved synteny-based genome zippers. This order provided an opportunity to step back and reappraise the overall extent of collinearity between barley and each of the three model grass genomes independently. Overall, 47, 20, and 33% of the loci anchored along the genome zippers were supported by conserved synteny in one, two, or all three model genomes, respectively. When barley gene order was compared with individual model genomes, we found that the number of conserved syntenic loci was similar in comparison with rice and sorghum (12,093 and 11,887, respectively) but was considerably higher with *Brachypodium* (14,422) reflecting a closer phylogenetic relationship. Overall, 20% of the loci anchored along the genome zippers were supported only by their order in the

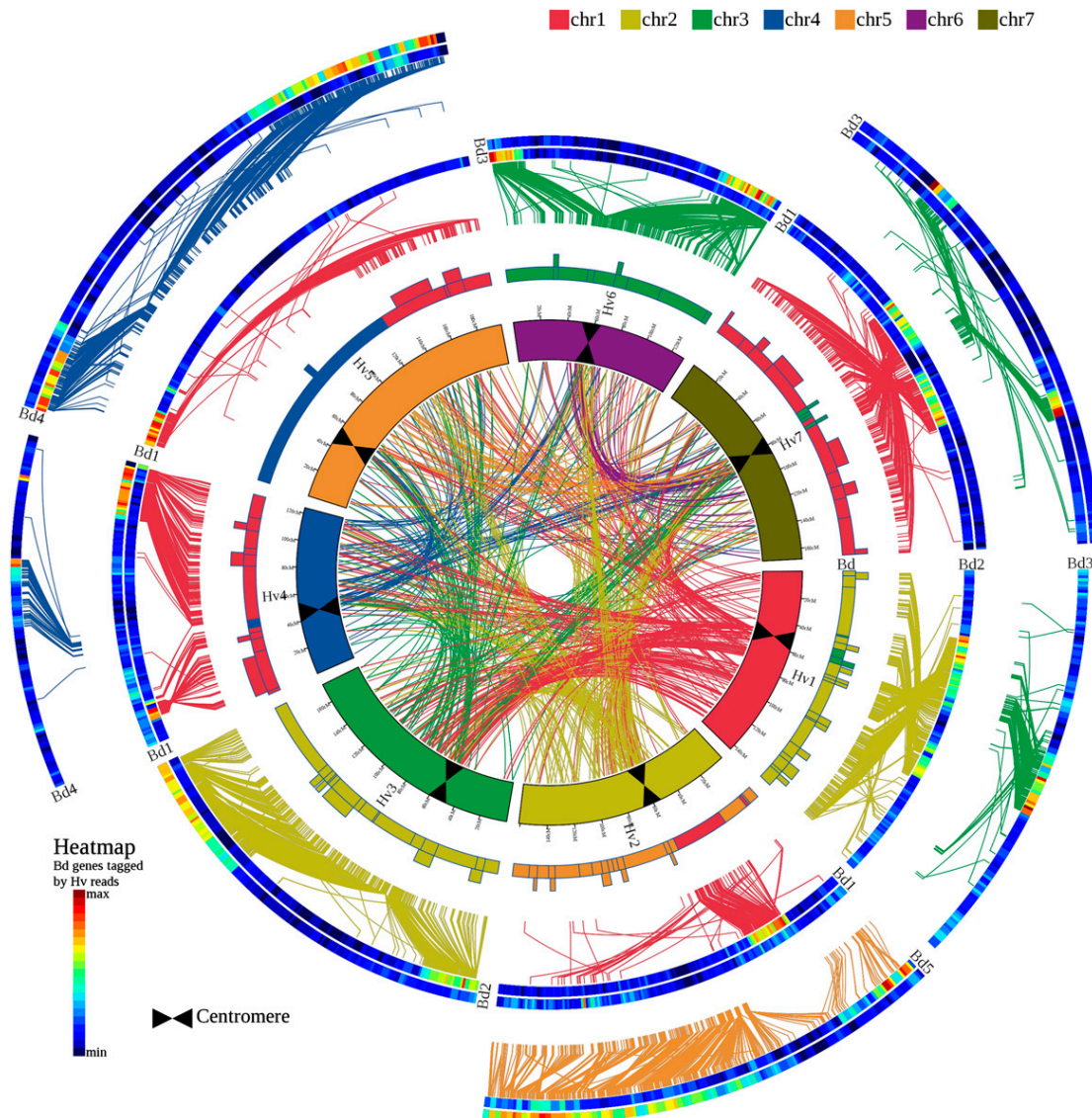


Figure 1. High-Resolution Comparative Analysis between Barley and *B. distachyon*.

High-density comparative analysis of the linear gene order of the barley genome zippers versus the sequenced model grass genome of *Brachypodium*. The figure includes four sets of concentric circles: the inner circle represents the seven chromosomes of barley scaled according to the barley genetic map (bars at 10-cM intervals). Each barley chromosome is assigned a color according to the sequence on the color key, starting with chr1 through chr7. The positions of the barley centromeres are indicated by black bars. Moving outwards, the second circle illustrates a schematic model of the seven barley chromosomes, but this time color-coded according to blocks of conserved synteny with the model genome. The color coding is again based on the sequence on the color key, but this time is based on the model genome linkage groups, starting with chr1 through chr5 for *Brachypodium*. Boxes extending from these colored bars indicate regions involved in larger-scale structural changes (e.g., inversions). The outer partially complete circles of heat map colored bars represent pseudomolecules of the model genome linkage groups arranged according to conserved synteny with barley 1H-7H. When pairs of adjacent heat map bars are shown, they illustrate where the homologs of a short (inner heat map bar) or a long (outer heat map bar) barley chromosome arm data set is allocated to the respective model genome pseudochromosome. The heat maps illustrate the density of genes hit by the 454 shotgun reads from the relevant barley chromosome arm. Conserved syntenic regions are highlighted by yellow-red-colored regions. Putative orthologs between barley and the model genomes are connected with lines (colored according to model genome chromosomes) between the second and third circles. Colored lines in the center represent putative paralogous relationships between barley chromosomes on the basis of fl-cDNA supported genes included in the genome zipper models of the seven barley chromosomes.

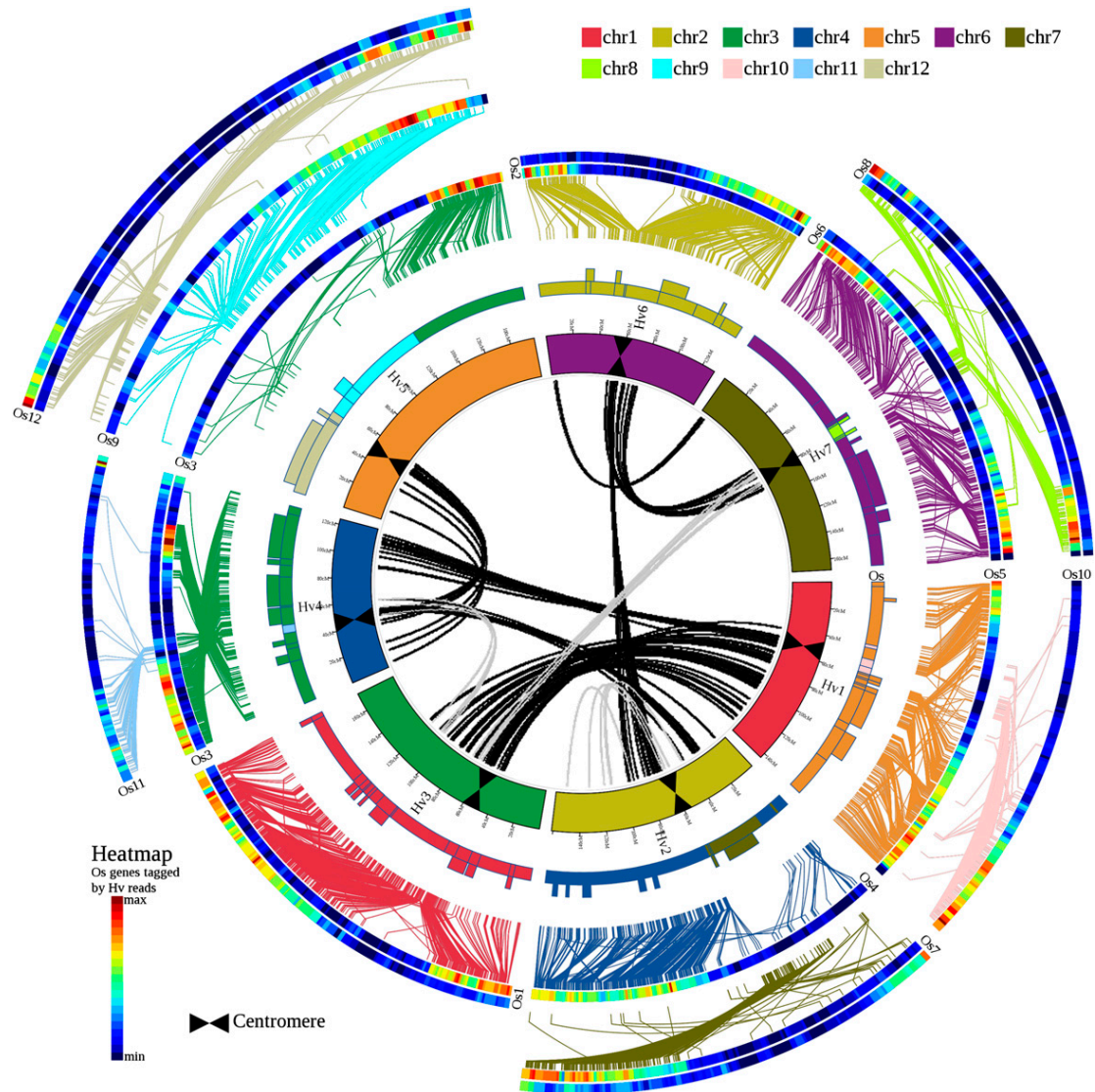


Figure 2. High-Resolution Comparative Analysis between Barley and Rice.

High-density comparative analysis of the linear gene order of the barley genome zippers versus the sequenced model grass genome of rice. Details are as provided in the Figure 1 legend. Putative orthologs between barley and the rice genomes are connected with lines (colored according to model genome chromosomes) between second, third, and fourth circles. In the center, nine major segmental duplications of the barley genome are visualized as statistically significant groups of paralogous genes. Each line represents a duplicated gene (paralogous gene pair). Black lines indicate ancestral duplications shared with the model grass genomes, and gray lines highlight barley-specific duplications.

Brachypodium genome, while 14.5 and 13% were exclusively supported by either rice or sorghum, respectively.

To reach the highest stringency and to reduce the risk of paralogous gene comparisons between species, we restricted all further steps of comparative genome analysis to genes incorporated in the genome zipper that had barley fl-cDNA support. Blocks of conserved synteny were apparent between barley and the model genomes, and these were consistent with previous observations among the different clades of grasses (Bolot et al., 2009) (Figures 1 to 3). Since the gene order in barley was guided by a dense genetic map, we first assigned and then

systematically compared the order and orientation of intervals among pairs or groups of genes to the model genomes. We identified numerous local inversions that appear to have either occurred specifically in barley, in one of the model genomes, or are shared between two genomes (Figure 3). For example, all inversions detected on the corresponding model genome segments of barley chromosome 3HL appear to be barley specific, since the order is conserved in all of the three model grass genomes. We then investigated patterns of ancestral whole-genome duplication in the barley genome. While this has been reported previously (Salse et al., 2009b; Thiel et al., 2009), the

Table 3. Genome Zipper Statistics: Genes, ESTs, and 454 Reads Associated with the Genome Zipper

Data Sets	1H MoBe	1H Morex	1H Betzes	2H	3H	4H	5H	6H	7H	All
Number of markers	332	332	332	468	445	314	492	337	397	2,785
Number of markers with associated gene from reference genome(s)	210	196	191	286	295	217	299	198	214	1,719
Number of matched array hybridization probes	732	n.d.	n.d.	2,044	1,502	1,242	1,935	1,407	2,003	10,865
Number of matched fl-cDNAs	1,676	1,287	1,247	1,619	1,628	1,255	1,474	1,058	1,395	10,105
Number of nonredundant sequence reads	51,972	28,485	17,716	29,250	30,576	21,402	25,262	19,536	22,420	200,418
Number of nonredundant ESTs	3,543	2,631	2,354	3,678	3,392	2,605	3,354	2,387	3,120	22,079
Number of <i>Brachypodium</i> genes	2,141	1,888	1,875	2,379	2,363	1,876	2,159	1,588	1,915	14,421
Number of rice genes	1,845	1,541	1,321	2,073	2,016	1,614	1,576	1,348	1,621	12,093
Number of sorghum genes	1,833	1,669	1,432	1,946	2,039	1,284	1,695	1,369	1,721	11,887
Number of nonredundant anchored gene loci in Genome Zipper	3,331	2,456	2,261	3,616	3,394	2,709	3,208	2,304	3,204	21,766

The table gives an overview of the data associated with and anchored along the chromosomal zippers. The number of markers is allocated to individual chromosomes. Data for the sequence collections of the individual cultivars used for 1H (Betzes and Morex) are listed separately as well as a combined data set (MoBe). n.d., not determined.

considerably increased gene coverage, particularly those with fl-cDNA support, along the genome zippers allowed us to recalculate paralogous relationships within the barley genome. This revealed a complex pattern of putatively duplicated genome segments (center of Figure 1). Using the alignment parameters and statistical tests defined by Salse et al. (2009a, 2009b), we identified nine major duplications (212 paralogous pairs) that cover 48% of the barley genome (center of Figure 2). Six of these corresponded to previously described ancestral segmental duplications shared between grass genomes. Three were considered barley specific. We thus substantiated in this analysis the previously reported paralogous gene content and duplicated block boundaries of such ancestral shared duplications in the Triticeae (Salse et al., 2008; Thiel et al., 2009).

There Is No Single Best Genomic Model for Barley

The principle uses of genomic models (certainly for wheat and barley) have been as predictors of regional candidate genes in positional cloning projects or for the development of gene-based markers that are tightly linked to a gene of interest. While these have been valid approaches, they frequently fail due to regional breakdown in the conservation of synteny. Given our newly available genomic information, we estimated the predictive value of individual model grass genomes for barley. We first associated the fl-cDNA supported linearly ordered barley genes with their orthologous counterparts in *Brachypodium*, rice, and sorghum. For this analysis, between 1247 and 1676 fl-cDNAs for each barley chromosome (average density of 9.3 fl-cDNAs per cM; 10,105 fl-cDNA/1090 cM) were tested. The extent of conserved synteny is not continuous for each barley genome segment/model genome species comparison. Therefore, a z-score within a sliding window (3-cM window, 0.1-cM shift) was calculated for comparison between each model species and barley to identify regions where conserved synteny was above or below average ($z > 0$ and $z < 0$, respectively) (Figure 3). Pronounced differences were observed along each chromosome, pinpointing regions where the degree of conserved synteny with individual model

genomes was greater than with others. These differences highlighted the advantage of adopting an integrative approach that used three model genomes in parallel to overcome limitations imposed by species-specific regional differences. It enabled us to anchor and order loci even in regions where one or two of the model genomes may have contained structural rearrangements, gene loss, or translocations.

Fast-Evolving Genes

All full-length coding sequences (fl-cDNAs) that were ordered and positioned in the genome zippers at conserved syntenic positions (10,105) were then used to calculate the ratio of nonsynonymous (K_a) to synonymous substitutions (K_s) against their orthologs in the respective model genomes. We calculated the K_a/K_s ratios for all compared genes. The K_a/K_s ratio measures the strength of selection acting on a protein sequence under the assumption that synonymous substitutions evolve neutrally. A ratio < 1 indicates purifying selection, and a ratio > 1 positive selection. The average K_a/K_s ratio of fl-cDNAs analyzed against *Brachypodium* (8160 genes), rice (7009 genes), and sorghum (6871 genes) is 0.21, 0.23, and 0.23, respectively, which indicates that the vast majority evolve under strong purifying selection. We chose a K_a/K_s ratio > 0.8 as a cutoff to identify rapidly evolving genes that includes genes with few evolutionary constraints or positively selected genes. In total, 105 barley genes exhibited K_a/K_s values > 0.8 in comparison to one (82 genes), two (15 genes), or all three (eight genes) model species, respectively (Figure 3; see Supplemental Figure 4 and Supplemental Data Set 9 online). These are assigned a wide range of putative molecular functions, including transcription factors and hormone responsive genes. Based on K_a/K_s ratios alone, these are candidates for conferring barley or Triticeae-specific phenotypic characteristics.

Rearrangements in Wheat A, B, and D Subgenomes

Within the Triticeae, the *Hordeum* (including barley) and the *Triticum* (including wheat) lineages split ~ 11 to 13 MYA (Gaut,

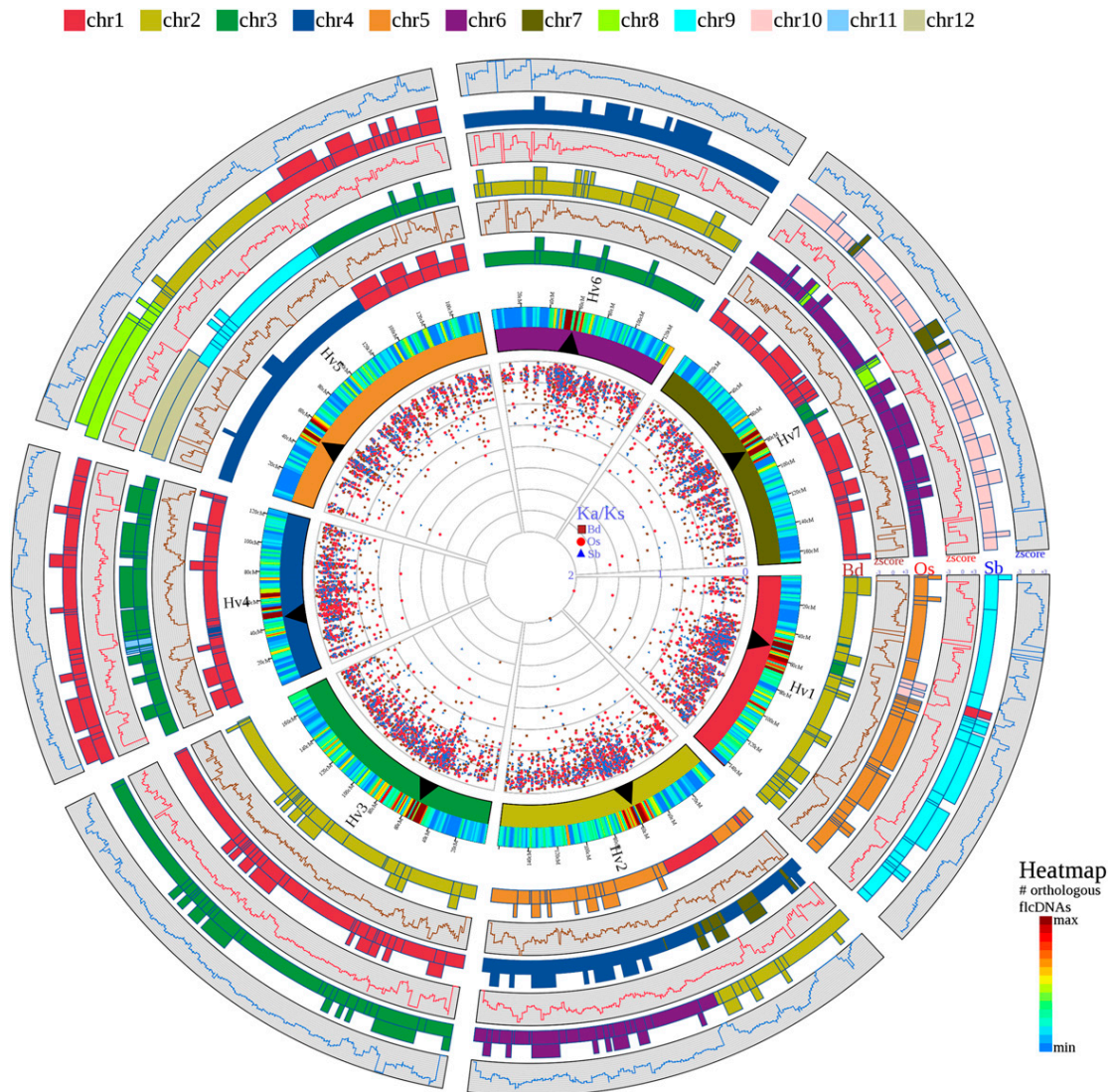


Figure 3. Barley-Centered Four-Genome Comparative View of Grass Genome Collinearity.

The seven barley chromosomes (Hv1 to Hv7) are depicted by the inner circle of colored bars exactly as in Figure 1. The heat map attached to each chromosome indicates the density of barley fl-cDNAs anchored and positioned along the chromosomes according to the genome zipper models. Gene density is colored according to the heat map scale. Moving outwards, the bars represent a schematic diagram of the barley chromosomes colored according to conserved synteny with the genomes of *Brachypodium* (Bd), rice (Os), and sorghum (Sb), respectively. In each case, the chromosome numbers and segments are colored according to the chromosome color code (i.e., chr1 through chr5 for Bd, chr1 through chr12 for Os, and chr1 through chr10 for Sb). As in Figure 1, boxes extending from the colored bars indicate structural changes (e.g., inversions) between the gene order in barley and the respective model genome. To the outside of each model genome chromosome, box graphs show the z-score derived from a sliding window analysis of the frequency of fl-cDNAs present at a conserved syntenic position with their corresponding orthologs in Bd, Os, and Sb, respectively (see Methods for a full description of the analysis). A z-score >0 indicates higher than the average conservation of synteny, and a z-score <0 highlights decreased syntenic conservation. The data points in the center of the diagram depict the K_a/K_s ratios between barley full-length genes and their orthologs in Bd, Os, and Sb. Values against Bd are plotted as dark red rectangles, against Os in red circles, and against Sb in blue triangles.

2002; Huang et al., 2002a), with the *Triticum* subgenomes radiating ~ 2.5 to 4.5 MYA. The tetraploid genome of *Triticum turgidum* (genome composition AABB) formed ~ 0.4 to 0.5 MYA, with a subsequent hybridization with *Aegilops tauschii* (DD) (~ 8000 years ago) forming the modern genome of allohexaploid bread wheat (genome composition AABBDD [Huang et al.,

2002b]). Using the genome zipper derived fl-cDNA gene indices assembled into pseudochromosomes, we tested the widely held view that barley (HH) contains an archetypal Triticeae genome by comparing it to the previously constructed high-density physical marker map of wheat (Qi et al., 2004) (Figure 4; see Supplemental Figure 5 online). As expected, most of the chromosome arms

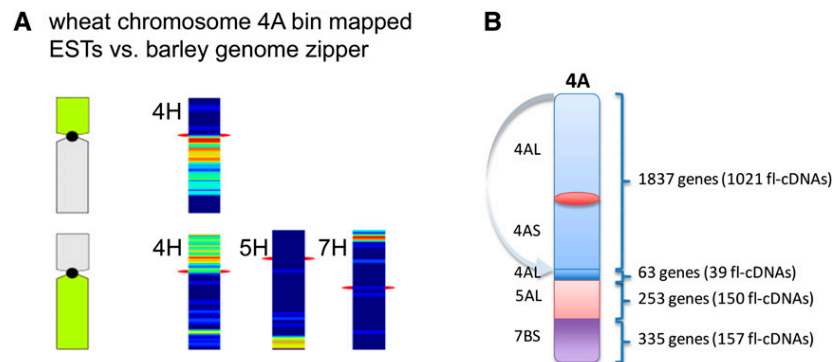


Figure 4. Structure of Wheat Chromosome 4A in Relation to the Barley Genome Zipper.

Wheat subgenome specific markers of chromosome 4A have been compared against the genome zipper chromosome model of barley (for a genome-wide overview, see Supplemental Figure 5 online). Orthologous regions are depicted and visualized by a heatmap.

(A) Wheat EST markers allocated to 4AS cross-match to barley genes on 4HL and markers allocated to 4AS, a small region on 4AL, 5AL, and 7BS cross-match to 4HL. Thus, a reciprocal translocation involving chromosomes 4A and 5A and a translocation from 7BS to 4AL was detected. Compared with barley 4H, wheat chromosome 4A contains a pericentromeric inversion.

(B) The barley genome zipper model allows the size of the affected regions to be estimated and the minimal number of genes located in these rearranged regions of the wheat chromosomes to be predicted.

exhibit well-conserved synteny with previously reported chromosomal translocations involving wheat 4A, 5A, and 7B accurately identified (Figure 4A; see Supplemental Figure 5 online). The availability of the barley genome zipper model allowed us also to estimate the gene content of the chromosomal fragments involved in such rearrangements (Figure 4B). Patterns of pericentric inversions could be deduced that confirmed previous observations involving wheat 2B, 3B, 4A, and 5A (Qi et al., 2006). The density of the compared data sets revealed regions that appear to be present in barley but lack counterparts in any of the homeologous wheat chromosomes (e.g., 1AS, 1AL, 2AL, and 2DL, all long arms of homeologous group 5 chromosomes; see Supplemental Figure 5 online); hence, blocks of barley genes cannot be assigned blocks of orthologs in the wheat bin map. Whether these regions have (1) been lost before the radiation of the wheat subgenomes, (2) have been integrated into barley independently, or (3) are simply not represented in the wheat EST bin map will only be resolved on the basis of more comprehensive data sets (e.g., by comparison to 454 sequence data of sorted wheat chromosomes). In addition, many small regions appeared to be absent in only one wheat subgenome, suggesting segmental loss possibly during or after major polyploidization events. Overall, at a structural level, no wheat subgenome was more similar to barley than any other and in terms of overall structural similarity and integrity, no conclusive evidence for more rapid structural evolution of any wheat subgenome was found. We conclude that most structural variation between A, B, and D genomes acts at a regional, maybe functional, level.

DISCUSSION

A complete reference genome sequence remains an aspiration for the barley research community, primarily due to technical and economic constraints resulting from the size and inherent com-

plexity of its 5.1-Gbp genome. As a step toward that goal, we report here a high resolution sequence-based gene map containing an estimated 86% of the genes in the barley genome. We present the genome as a set of seven genome zippers that embrace the well-established conservation of synteny shown to exist among grass genomes. We propose that these genome zippers provide a high utility surrogate for both the barley genome itself and for closely related Triticeae cereals and are a high-resolution infrastructure upon which structural genomic information, such as physical maps, can be superimposed (Schulte et al., 2009).

The data used to derive the genome zippers were generated from low-pass 454 shotgun sequencing of individual flow-sorted barley chromosome/chromosome arm preparations and hybridization of equivalent subgenomic DNA preparations against a barley long oligonucleotide (gene) array. Both data sets are independent, exhibit high sensitivity and specificity, and show excellent concordance (>95%). Combining a recently developed 2785 gene-based genetic marker map (Close et al., 2009) with synteny information from model grass genomes provided the framework that enabled us to produce a highly structured and ordered sequence-based map comprising of 21,766 ordered barley genes. We consider that this ordering of genes along the chromosomes has reached a density and precision that can only be exceeded by a complete barley genome sequence.

This high-resolution view of the barley genome illuminates issues that have been faced in cereal genetics and breeding for many years. For example, we observed that 3125 genes fall into regions of the genome classified as genetic centromeres. These are regions where gene order cannot be established by meiotic mapping and where even crude assignment of genes to either proximal or distal chromosome arms has previously proved impossible. We were not only able to assign all but nine of these 3125 genes to the proximal or distal arms but also to propose a linear order. This allowed us to undertake genome scale analyses that included a fine-detail reappraisal of conservation of synteny

with sequenced grass genomes, including an assessment of regional variation in the degree of conservation, an exploration of large-scale ancestral duplications, rearrangements, and more recent and local duplications. We present these for immediate exploitation by the Triticeae genetics and genomics community for both fundamental (i.e., physical map anchoring) or applied (i.e., candidate gene identification) purposes.

The clustering of genes toward genetic centromeres of barley has been well documented (Stein et al., 2007). In this study, one-third of all genes (6788 genes) in the genome zippers are located within 10-cM intervals that encompass each genetic centromere (6.4% of the entire barley genetic map). In wheat, sequencing megabase-sized BAC contigs selected from distributed regions of the chromosome 3B physical map revealed the presence of genes throughout the physical length of the chromosome, with a twofold higher concentration toward the telomeres (Choulet et al., 2010). Since regions with low recombination frequency per physical unit (hence, the regions around genetic centromeres) may extend in barley over as much as half a barley chromosome (Künzel et al., 2000), it can be expected that gene distribution in barley will follow a similar pattern as observed for wheat chromosome 3B. Unfortunately, this will place severe constraints on positional gene isolation for as many as one-third of barley genes. While the genome zippers will still provide a rich source of information for gene-based marker development and candidate gene identification in these regions, it is likely that innovative genetic strategies, such as deletion mapping or genome-wide association studies in highly diverse (e.g., wild) populations that have had orders of magnitude more opportunity for recombination, may be required (Waugh et al., 2009).

Due to their close evolutionary relationship, we investigated the degree of structural conservation between barley and wheat in more detail. As reported previously by comparing transcript map data to sequenced model genomes (Bolot et al., 2009), at a global level, a high degree of similarity was confirmed between the two species. Wheat chromosome 4A represents a notable exception, being a highly rearranged chromosome involving a large-scale inversion and two interchromosomal translocations (Mickelson-Young et al., 1995; Nelson et al., 1995; Miftahudin et al., 2004). The novelty of comparing the genome zipper model of barley to the wheat EST deletion bin map is that a better estimate of the genes involved can be made than by comparison to more distantly related models. Thus, several centromeric inversions that have been reported for the wheat genome (Qi et al., 2006) could also be deduced from our high-density comparison. These rearrangements appear to be wheat specific, not occurring at this frequency in the diploid barley genome. An apparent pericentromeric inversion shared by all wheat group one chromosomes likely indicates that the inversion occurred in barley in the period between the separation of the barley lineage and the radiation of wheat (i.e., some 11 to 4.5 to 2.5 MYA). Confirming this will require further experimentation. Based on the resolution of the bin-mapped wheat EST markers, many small regions appear to be missing from the individual wheat subgenomes. In contrast with all previous comparative analyses in the Triticeae, the genome zippers allow both the genetic size and the conserved (syntenic) gene content of the affected regions to be determined.

On a structural basis, none of the individual wheat A, B, or D subgenomes was more closely or distantly related to the H genome with numerous variations apparent in only one or two wheat subgenomes. This implies a highly complex, mosaic type, structural evolution of the A, B, and D subgenomes after radiation and the two subsequent polyploidization events that lead to the genomic composition of modern wheat (AABBDD). Such an outcome may have been predicted as a consequence of profound changes in genome structure and function induced by genomic shock in the early generations following the development of the allopolyploid (Chen, 2007). Indeed, in newly formed synthetic wheats, the reproducible elimination of specific sequences accounting for up to ~14% of the genomic DNA has been demonstrated and proposed to provide a physical mechanism for genetic diploidization in new allopolyploids (Feldman et al., 1997; Ozkan et al., 2001; Shaked et al., 2001). While local rearrangements, expansions, and single gene loss is beyond the currently available resolution, once a more complete genome sequence is available, the evolutionary dynamics between the H genome and the A, B, and D genomes of wheat can be expected to give important insights into genomic evolution and the structural and functional consequences of allopolyploidization.

We estimate that the barley genome contains in the order of 32,000 genes. Our estimate was based on (1) a stringent comparison of a comprehensive set of barley fl-cDNAs against sequenced model grass genomes and (2) the number of genes detected in 454 sequence and array-based data obtained from sorted barley chromosomes that matched a model genome homolog. Comparisons against model genomes detected 21,240 nonredundant genes. Given a sensitivity of 0.86, this would scale to 24,700 barley genes with a sequence homolog for the complete genome. Analysis of a set of 23,588 nonredundant barley fl-cDNAs revealed that using our stringent criteria 23% lack a sequence homologous counterpart in the model genomes. Taking this observation into account, we expect ~32,000 genes to be present in the barley genome. This number is remarkably consistent with gene number estimates for diploid grass model genomes (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; The International Brachypodium Initiative, 2010).

An estimate of 50,000 genes was given for a diploid wheat genome on the basis of megabase-sized BAC contig sequencing of chromosome 3B and short-read (Illumina/Solexa) survey sequencing of sorted 3B chromosomes (Choulet et al., 2010). Since the approaches used and the underlying sequence data differ, our analysis is not directly comparable to that of wheat 3B. For example, analysis of closely related expanded gene families, such as locally duplicated genes or translocated duplicated genes, cannot be appropriately addressed in shotgun sequences. Thus, paralogous gene families might in part have been interpreted as single genes, and consequently our gene number estimate may represent a lower limit.

The barley fl-cDNAs at conserved positions in all four genomes in the genome zipper allowed us to conduct a global survey for fast-evolving genes in barley by comparison to one, two, or all three sequenced model grass genomes and identified 105 genes with significant K_a/K_s values. We identified only eight barley genes that exhibited K_a/K_s ratios >0.8 in comparison to all three

model grass genomes. Three genes were of unknown function and the remaining five genes can all be assigned to developmental roles based on their annotation. Two are transcription factors: one (NIASHv2057H16; see Supplemental Data Set 9 online) exhibiting strong similarity to a homeobox transcription factor *Oshox24* (Agalou et al., 2008), which in rice shows differential expression in roots and panicle tissues at maturation. One was a rapid alkalization factor, a class of genes shown to be involved in root and maybe also pollen development in different plant species (Germain et al., 2005; Wu et al., 2007; Zhang et al., 2010). Two genes encode homologs of pectin-methylesterase inhibitors (PMEIs). PMEIs inhibit the enzyme pectin-methylesterase, which is required for demethoxylation of methylated pectins, a necessary step before degradation by pectin-depolymerizing enzymes. Pectin-methylesterases are ubiquitous enzymes in plants and their fine-tuned regulation (i.e., by PMEI) may be crucial during steps of development that require cell wall modifications (for review, see Jolie et al., 2010). It is tempting to speculate about the possible role of these five genes in specific developmental processes in barley. However, the significance of our observations as well as other possible mechanisms leading to evolution of species- and clade-specific traits like diversification of gene expression regulation (reviewed in Rosin and Kramer, 2009) will require future experimental testing.

Linear gene order information as provided by the barley genome zippers will be vital for the generation of a complete genome reference for barley. The development of a high information content fingerprint BAC-based physical map of the barley genome is well advanced (Schulte et al., 2009), and this effort will likely profit from the presented data sets for anchoring the physical map to a genetic/syntenic framework. Referring to the model character of barley for other Triticeae genomes, such a detailed barley framework will play a pivotal role in the assembly of data that could be generated for other Triticeae species. An obvious primary target is of course wheat (Kubaláková et al., 2002) and survey sequencing of chromosomes for the construction of a genome-wide collection of wheat genome zippers has already been initiated (IWGSC; <http://www.wheatgenome.org/Projects>). The approach is equally attractive for rye (*Secale cereale*; Kubaláková et al., 2003). More generally, the approach may be adopted as an economic and technical paradigm for other unsequenced orphan crop genomes where individual chromosomes, chromosome arms, or translocations can be separated by flow sorting techniques. These include legumes such as chickpea (*Cicer arietinum*; Vlácilová et al., 2002), garden pea (*Pisum sativum*; Neumann et al., 2002), and field bean (*Phaseolus vulgaris*; Doležel and Lucretti, 1995) where the feasibility of chromosome flow sorting has previously been demonstrated.

The genome zipper-based linear gene order model of two-thirds of all barley genes will open a path toward contextualized genome-wide diversity analysis in barley. Currently available NGS technology allows for whole-genome shotgun sequencing and de novo assembly to draft sequence quality even of complex mammalian genomes (Li et al., 2010). With the currently available technology, a similar attempt in barley could lead to assembled gene sequence information and thus provide a genomic reference for genes of the genome zipper. Using this information as reference for resequencing, polymorphism surveys will become

a realistic endeavor for the majority of the barley gene space. In combination with the appropriate plant material, such as the well-characterized mutant collections available in barley (Druka et al., 2010), we may soon be able to clone the genes that are responsible for many phenotypic traits by direct resequencing, similar to approaches successfully applied in *Arabidopsis thaliana* (Schneeberger et al., 2009).

METHODS

Purification and Amplification of Chromosomal DNA

Intact mitotic chromosomes/arms were isolated by flow cytometric sorting from barley *Hordeum vulgare* cultivar Morex and cv Betzes (1H) and wheat (*Triticum aestivum*)-barley telosome addition lines (2HS-7HL arms originating from cv Betzes). The purity in the sorted fractions was determined by fluorescence in situ hybridization essentially as described previously (Suchánková et al., 2006). The DNA of sorted chromosomes was purified and amplified by MDA as described previously (Šimková et al., 2008).

Roche 454 Sequencing

DNA amplified from sorted chromosomes was used for 454 shotgun sequencing. Five micrograms of individual chromosome arm MDA DNAs were used to prepare the 454 sequencing libraries using the GS Titanium General Library preparation kit following the manufacturer's instructions (Roche Diagnostics). The 454 sequencing libraries were processed using the GS FLX Titanium LV emPCR (Lib-L) and GS FLX Titanium Sequencing (XLR70) kits (Roche Diagnostics) according to the manufacturer's instructions. Sequencing details are summarized in Table 1 and Supplemental Table 1 online.

Microarray Construction and Analysis

A custom microarray SCRI_Hv35_44k_v1 (Agilent design 020599) representing 42,302 barley sequences was generated. Barley sequences for this design were selected from a total of 50,938 unigenes from HarVest assembly 35 (<http://www.harvest-web.org/>) representing ~450,000 ESTs. Selection criteria were based upon the ability to define orientation derived from (1) homology to members of the nonredundant protein database (NCBI nr), (2) homology to ESTs known to originate from directional cDNA libraries, and (3) presence of a significant poly(A) tract. The microarray was designed with one 60mer probe per selected unigene in 4 × 44k format using default parameters in the Web-based Agilent eArray software (<https://earray.chem.agilent.com/earray/>) and includes recommended QC control probes. Full details of array design, probe sequences, and unigene accession numbers can be found at Array-Express (<http://www.ebi.ac.uk/microarray-as/ae/>; accession number A-MEXP-1728). Due to the redundancy in the EST-based unigene data set used as a basis for array design, the microarray comprised an estimated 25 to 32,000 nonredundant barley genes (Michael Bayer, personal communication; each gene was represented on average by ~1.3 to 1.7 probes per genes).

Fluorescent Labeling of Chromosome DNA and Hybridization to Barley Microarrays

Amplified chromosomal DNA was labeled using a modified Bioprime DNA labeling system (Invitrogen). For each sample, 2 μg amplified genomic DNA in 21 μL was added to 20 μL Random Primer Reaction Buffer and denatured at 95°C for 5 min prior to cooling on ice. To this, 5 μL modified

10× deoxynucleotide triphosphate mix (1.2 mM each of dATP, dGTP, and dTTP, 0.6 mM dCTP, 10 mM Tris, pH 8.0, and 1 mM EDTA), 3 μL of either Cy3 or Cy5 dCTP (1 mM), and 1 μL Klenow enzyme was added and incubated for 16 h at 37°C. Labeled samples for each array were combined and unincorporated dyes removed using the MinElute PCR purification kit (Qiagen) as recommended, eluting twice with 1× 10 μL sterile water. Specific activities of incorporated dyes (nmol/μg DNA) were estimated using spectrophotometry.

The design of the microarray experiment is detailed in ArrayExpress (accession number E-TABM-1063) and ensured that independent replicate samples of each amplified chromosome arm were labeled once with each of two fluorescent dyes, Cy3 and Cy5, to minimize dye bias. Microarray hybridization and washing were conducted according to the manufacturer's protocols as for gene expression arrays (Agilent Two-Color Microarray-Based Gene Expression Analysis, version 5.5). For each array, 20 μL purified labeled samples were added to 5 μL 10× blocking agent and heat denatured at 98°C for 3 min then cooled to room temperature. GE Hybridization Buffer HI-RPM (25 μL) was added and mixed prior to hybridization at 65°C for 17 h at 10 rpm. Array slides were dismantled in Agilent Wash 1 buffer and washed in Wash 1 buffer for 1 min, then Agilent Wash 2 buffer for 1 min, and centrifuged dry. Hybridized slides were scanned using an Agilent G2505B scanner at resolution of 5 μm at 532 nm (Cy3) and 633 nm (Cy5) wavelengths with extended dynamic range (laser settings at 100 and 10%).

Microarray Data Extraction and Analysis

Microarray images were imported into Agilent Feature Extraction (FE v.10.5.1.1) software and aligned with the appropriate array grid template file (020599_D_F_20080612). Intensity data and QC metrics were extracted using a suitable FE protocol (GE2-v5_95_Feb07), and data from each array were normalized in FE using the LOWESS (locally weighted polynomial regression) algorithm to minimize differences in dye incorporation efficiency (Yang et al., 2002). Entire normalized data sets for both channels of each array were loaded into GeneSpring (v.7.3.1) software for further analysis. Data were subjected to additional normalization whereby values were set to a minimum of 5.0, data from each array were scaled to the 50th percentile of all measurements on the array, and the signal from each probe was subsequently normalized to the median of its values. Unreliable data with consistently low probe intensity levels (raw values <100) in all replicate samples were discarded. Statistical filtering of data for each experiment was performed using analysis of variance with Benjamini and Hochberg (Benjamini and Hochberg, 1995) false discovery rate for multiple testing correction (P value <0.005). Heat maps were generated from filtered probe/gene lists using an average linkage clustering algorithm based upon Pearson correlation using default parameters in GeneSpring. Clustered probes enriched for each chromosome arm were selected manually from the gene tree.

General Sequence Analysis

Repeat Masking of 454 Sequence Data

To determine genic regions covered by 454 sequencing data, the content of repetitive DNA per sequence read was masked after being identified using Vmatch (<http://www.vmatch.de>) against the MIPS-REdat Poaceae v8.2 repeat library (contains known grass transposons from the Triticeae Repeat Database, <http://wheat.pw.usda.gov/ITMI/Repeats>, as well as de novo detected LTR retrotransposon sequences from several grass species, specifically, maize [*Zea mays*], 12434; sorghum [*Sorghum bicolor*], 7500; rice [*Oryza sativa*], 1928; *Brachypodium distachyon*, 466; wheat, 356; and barley, 86 sequences) by applying the following parameters: 60% identity cutoff, 30-bp minimal length, seed length 14, exdrop 5, and e-value 0.001.

Identification of Genetic Markers in the 1H-7H Data Sets

The repeat-masked sequence collections from all seven barley chromosomes were compared (BLASTN) against 2785 nonredundant (of total 2943) EST-based markers (Close et al., 2009; <http://harvest.ucr.edu>) under optimized parameters (-r 1 -q -1 -W 9 -G 1 -E 2: -r reward for a nucleotide match, default = 1; -q penalty for a nucleotide mismatch, default = -3; -W word size, default; -G cost to open a gap, default = -1; -E cost to extend a gap, default = -1). Only BLAST matches exceeding an identity threshold of 98% and an alignment length of 50 bp were considered.

A Nonredundant Set of Barley fl-cDNA

In this study, a set of 5006 (Sato et al., 2009b) and a set of 23,623 barley full-length cDNAs (Matsumoto et al., 2011) was used for sequence comparison. All redundant cDNA sequences were removed and a database of 23,588 nonredundant fl-cDNAs was generated for further steps of analysis using CD-HIT-EST (<http://www.bioinformatics.org/cd-hit/>) applying the following parameter settings: -c 0.98 and -n 8 (-c sequence identity threshold, default 0.9; -n word length, default 5).

Overall Gene Content in the Combined Chromosome-Specific Barley Sequence Data Set

To estimate the number of barley genes that have been captured in the barley sequence collection generated by Roche 454 sequencing, BLASTX (Altschul et al., 1990) comparisons were performed with the repeat-filtered 454 sequence reads, the microarray probe sets, and the nonredundant fl-cDNAs against *Brachypodium*, rice, and sorghum proteins (*Brachypodium* genome annotation v1.2 [<ftp://ftpmips.helmholtz-muenchen.de/plants/brachypodium/v1.2>]; rice RAP-DB genome build 4 [<http://rapdb.dna.arc.go.jp>]; sorghum genome annotation v1.4 [<http://genome.jgi-psf.org/Sorbi1/Sorbi1.download.ftp.html>]; Paterson et al., 2009). The number of tagged genes and the number of gene matching reads and fl-cDNAs were counted after filtering according to the following criteria: (1) the best hit display with a similarity >75% and (2) an alignment length ≥30 amino acids. To increase specificity, microarray probes (length of 60 nucleotides) were associated with their respective cognate EST. These were used for subsequent integration using the parameters above.

Association of Barley fl-cDNA and EST to Individual Barley Chromosomes (Arms)

The putative chromosomal origin of barley cDNA and EST collections (HarvEST barley v1.73, assembly 35; <http://harvest.ucr.edu/>) was determined by BLASTN comparison against the repeat masked shotgun sequence reads from all seven barley chromosomes. Only the best hits with an identity of >98% and a minimal alignment length of 50 bp were considered. Each cDNA or EST was assigned to a particular chromosome (arm) if at least 80% of associated shotgun sequence reads were assigned to the same chromosome.

Assessment of Linear Gene Order in Barley (Genome Zipper)

Conserved synteny between three model grass genomes was used as a template to develop a linear gene order model (genome zipper) of the genes assigned to individual barley chromosomes by the analysis steps described above. The workflow toward a so-called genome zipper of a given barley chromosome was designed to structure and order barley genes identified either by 454 shotgun sequencing or microarray hybridization to sorted chromosomal DNA on the basis of collinearity to

model grass genomes. As a first step, the repeat masked shotgun sequences and array probes associated with each individual chromosome/chromosome arm were compared (BLASTX) against the three reference genomes *Brachypodium*, sorghum, and rice. Genes from syntenic regions, as defined by the density of homology matches, from the three genomes were selected and compared with the dense gene-based marker map of barley, which served as a scaffold to anchor collinear segments from model genomes. This step was performed for the three model grass genomes and results are interlaced based on joint marker associations as well as best bidirectional hit (bbh) classification. Sequence-tagged genes are anchored to the marker scaffold and additional tagged genes without barley marker association were ordered following the concept of conserved synteny and closest evolutionary distance. Finally the integrated syntenic scaffolds were associated with fl-cDNAs, array probes, ESTs, and shotgun reads that exhibited matches to the syntenic genes and the barley EST-based marker. Genome zipper-based tentative gene order, including associated information, is provided in Supplemental Data Sets 2 to 8 online.

Analysis of Conserved Synteny

The degree of conserved synteny against each of the model grass genomes rice, sorghum, and *Brachypodium* was calculated using a sliding window approach. For each genetic position (3-cM window, window shift 0.1 cM), the number of syntenic genes (classified as syn+) divided by the sum of all genes (syntenic and nonsyntenic, syn+ and syn-) was calculated (=conserved synteny). Genome-wide local differences were analyzed by calculating the z-score to indicate regions with above average and below average conservation ($z > 0$ and $z < 0$, respectively).

Calculation of Synonymous and Nonsynonymous (K_a/K_s) Substitution Rates

Sequence divergence as well as speciation event dating analysis based on the rate of nonsynonymous (K_a) versus synonymous (K_s) substitutions was calculated using the YY00 program within the PAML suite (phylogenetic analysis by maximum likelihood) (Nei and Gojobori, 1986; Yang, 2007). Only high-quality alignments and depending on the number of detectable orthologs 2, 3, or 4 sequences were used.

Analysis of Traces of Genome Duplications in Barley

Analysis was performed using the procedure and definitions defined previously (Salse et al., 2009a, 2009b) as well as by a best BLAST hit (bbh) strategy. Sequence divergence and speciation event dating analysis based on the rate of nonsynonymous (K_a) versus synonymous (K_s) substitutions was calculated and an average substitution rate (r) of 6.5×10^{-9} substitutions per synonymous site per year (Gaut et al., 1996; SanMiguel et al., 1998). The time (T) since gene insertion has been estimated using the formula $T = K_a/r$.

Analysis of Synteny between Barley and Homoeologous Wheat Chromosomes

Barley fl-cDNAs integrated in the barley genome zipper were concatenated following the order assigned in the genome zipper (with spacer sequences between individual genes) to result in approximated chromosome scaffolds. These scaffolds were compared against the high-density physical wheat transcript map (deletion bin map; Qi et al., 2004) using BLASTN (identity $\geq 85\%$, match length ≥ 100 nucleotides). Matching and nonmatching genes were depicted independently for the A, B, and D derived markers in a heat map following the assigned gene order from the barley genome zippers.

Data Availability and Accession Numbers

The nonredundant set of 23,588 fl-cDNAs was generated from a set of 5006 fl-cDNAs (Sato et al., 2009b; accession numbers AK248134 to AK253139) and a set of 23,623 fl-cDNAs (Matsumoto et al., 2011; accession numbers AK353559 to AK377172). All 454 sequence information in this study generated from flow-sorted chromosomes was submitted to the European Bioinformatics Institute sequence read archive under accession number ERP000445. A database for sequence homology search (BLAST) is provided at <http://webblast.ipk-gatersleben.de/barley/>. All data contained in the genome zipper models can be downloaded as Excel spread sheets from <http://mips.helmholtz-muenchen.de/plant/triticeae/genomes/index.jsp>.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Hierarchical Clustering of Microarray Hybridization to Sorted Chromosomal DNA of Barley.

Supplemental Figure 2. Flow Chart for the Genome Zipper Analysis Pipeline.

Supplemental Figure 3. Conservation of Synteny between Barley and Sorghum.

Supplemental Figure 4. Number of Genes with K_a/K_s Values of >0.8 between Barley and *Brachypodium*, Rice, and Sorghum.

Supplemental Figure 5. Global Analysis of Barley/Wheat Conserved Synteny on the Basis of the Genome Zipper Model.

Supplemental Table 1. Sequencing Statistics for Individual Chromosomes and Chromosome Arm.

Supplemental Table 2. Accuracy (the Proportion of True Results) of Sequence Read Distribution to Mapped Barley Markers.

Supplemental Table 3. Summary of Flow-Sorted Chromosome Fractions and Their Purities as Determined by FISH.

Supplemental Data Set 1. 454 Sequence Read Distribution to Barley EST-Based Markers.

Supplemental Data Sets 2 to 8. Genome Zipper of Barley Chromosomes 1H to 7H, Respectively.

Supplemental Data Set 9. Genes with Evidence for Positive Selection as Based on K_a/K_s Signatures.

ACKNOWLEDGMENTS

We thank Jarmila Čihalíková, Romana Šperková, and Zdenka Dubská for assistance with chromosome sorting and DNA amplification as well as Susanne König and Jana Schellwat for Roche 454 sequencing of sorted chromosomes. We are grateful to the Cereal Research Institute Kroměříž for the seeds of barley cultivar Betzes and the National Bioresource Project-Wheat (Japan) for seeds of wheat-barley telosome addition lines. We thank David Marshall, Karl Schmid, and three anonymous reviewers for constructive and helpful comments on the manuscript. We kindly acknowledge Bjoern Usadel, Birgit Kersten, Diego Riano-Pachon, and Doreen Pahlke from GABI-PD (Genome Analysis in the Biological system plant-Primary Database) for support with submission of sequence data sets to European Bioinformatics Institute. This work was financially supported by the following grants: 0314000 GABI Barlex from the Bundesministerium für Bildung und Forschung to N.S., K.F.X.M., M.P., and U.S.; FP7-212019 TriticeaeGenome from the European Union commission to N.S., R.W., K.F.X.M., and J.D.; and the Ministry of Education, Youth, and Sports of the Czech Republic and the European Regional Development Fund (Operational Programme

Research and Development for Innovations No. CZ.1.05/2.1.00/01.0007. J.D., R.W., K.F.X.M., and N.S. collaborated in the frame of European Cooperation in Science and Technology action FA0604 Tritigen.

Received December 20, 2010; revised March 10, 2011; accepted March 18, 2011; published April 5, 2011.

REFERENCES

- Abrouk, M., Murat, F., Pont, C., Messing, J., Jackson, S., Faraut, T., Tannier, E., Plomion, C., Cooke, R., Feuillet, C., and Salse, J.** (2010). Palaeogenomics of plants: Synteny-based modelling of extinct ancestors. *Trends Plant Sci.* **15**: 479–487.
- Agalou, A., et al.** (2008). A genome-wide survey of HD-Zip genes in rice and analysis of drought-responsive family members. *Plant Mol. Biol.* **66**: 87–103.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J.** (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- Beales, J., Turner, A., Griffiths, S., Snape, J.W., and Laurie, D.A.** (2007). A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **115**: 721–733.
- Benjamini, Y., and Hochberg, Y.** (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**: 289–300.
- Bennett, M.D., and Smith, J.B.** (1976). Nuclear DNA amounts in angiosperms. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **274**: 227–274.
- Bolot, S., Abrouk, M., Masood-Quraishi, U., Stein, N., Messing, J., Feuillet, C., and Salse, J.** (2009). The 'inner circle' of the cereal genomes. *Curr. Opin. Plant Biol.* **12**: 119–125.
- Chen, Z.J.** (2007). Genetic and epigenetic mechanisms for gene expression and phenotypic variation in plant polyploids. *Annu. Rev. Plant Biol.* **58**: 377–406.
- Choulet, F., et al.** (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell* **22**: 1686–1701.
- Close, T.J., et al.** (2009). Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* **10**: 582.
- Close, T.J., Wanamaker, S.I., Caldo, R.A., Turner, S.M., Ashlock, D.A., Dickerson, J.A., Wing, R.A., Muehlbauer, G.J., Kleinhofs, A., and Wise, R.P.** (2004). A new resource for cereal genomics: 22K barley GeneChip comes of age. *Plant Physiol.* **134**: 960–968.
- Devos, K.M.** (2005). Updating the 'crop circle'. *Curr. Opin. Plant Biol.* **8**: 155–162.
- Doležel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysák, M.A., Nardi, L., and Obermayer, R.** (1998). Plant genome size estimation by flow cytometry: Inter-laboratory comparison. *Ann. Bot. (Lond.)* **82**: 17–26.
- Doležel, J., and Lucretti, S.** (1995). High-resolution flow karyotyping and chromosome sorting in *Vicia faba* lines with standard and reconstructed karyotypes. *Theor. Appl. Genet.* **90**: 797–802.
- Druka, A., Franckowiak, J., Lundqvist, U., Bonar, N., Alexander, J., Houston, K., Radovic, S., Shahinnia, F., Vendramin, V., Morgante, M., Stein, N., and Waugh, R.** (2010). Genetic dissection of barley morphology and development. *Plant Physiol.* **155**: 617–627.
- Druka, A., et al.** (2006). An atlas of gene expression from seed to seed through barley development. *Funct. Integr. Genomics* **6**: 202–211.
- Feldman, M., Liu, B., Segal, G., Abbo, S., Levy, A.A., and Vega, J.M.** (1997). Rapid elimination of low-copy DNA sequences in polyploid wheat: A possible mechanism for differentiation of homoeologous chromosomes. *Genetics* **147**: 1381–1387.
- Fu, D., Szűcs, P., Yan, L., Helguera, M., Skinner, J.S., von Zitzewitz, J., Hayes, P.M., and Dubcovsky, J.** (2005). Large deletions within the first intron in *VRN-1* are associated with spring growth habit in barley and wheat. *Mol. Genet. Genomics* **273**: 54–65.
- Gaut, B.S.** (2002). Evolutionary dynamics of grass genomes. *New Phytol.* **154**: 15–28.
- Gaut, B.S., Morton, B.R., McCaig, B.C., and Clegg, M.T.** (1996). Substitution rate comparisons between grasses and palms: Synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. *Proc. Natl. Acad. Sci. USA* **93**: 10274–10279.
- Germain, H., Chevalier, É., Caron, S., and Matton, D.P.** (2005). Characterization of five RALF-like genes from *Solanum chacoense* provides support for a developmental role in plants. *Planta* **220**: 447–454.
- Huang, S., Sirikhachornkit, A., Faris, J.D., Su, X., Gill, B.S., Haselkorn, R., and Gornicki, P.** (2002a). Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Mol. Biol.* **48**: 805–820.
- Huang, S., Sirikhachornkit, A., Su, X., Faris, J., Gill, B., Haselkorn, R., and Gornicki, P.** (2002b). Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc. Natl. Acad. Sci. USA* **99**: 8133–8138.
- International Brachypodium Initiative** (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**: 763–768.
- International Rice Genome Sequencing Project** (2005). The map-based sequence of the rice genome. *Nature* **436**: 793–800.
- Jolie, R.P., Duvetter, T., Van Loey, A.M., and Hendrickx, M.E.** (2010). Pectin methylesterase and its proteinaceous inhibitor: A review. *Carbohydr. Res.* **345**: 2583–2595.
- Jordan, T., Seeholzer, S., Schwizer, S., and Keller, B.** (2010). The wheat *Mla* homolog *TmMla1* exhibits an evolutionary conserved function against powdery mildew in both wheat and barley. *Plant J.* **65**: 610–621.
- Kubaláková, M., Vrána, J., Čihalíková, J., Simková, H., and Doležel, J.** (2002). Flow karyotyping and chromosome sorting in bread wheat (*Triticum aestivum* L.). *Theor. Appl. Genet.* **104**: 1362–1372.
- Kubaláková, M., Valárik, M., Barto, J., Vrána, J., Čihalíková, J., Molnár-Láng, M., and Doležel, J.** (2003). Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. *Genome* **46**: 893–905.
- Künzel, G., Korzun, L., and Meister, A.** (2000). Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* **154**: 397–412.
- Lander, E.S., and Waterman, M.S.** (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**: 231–239.
- Li, R., et al.** (2010). The sequence and de novo assembly of the giant panda genome. *Nature* **463**: 311–317.
- Matsumoto, T., et al.** (2011). Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from twelve clone libraries. *Plant Physiol.* **156**: 20–28.
- Mayer, K.F.X., et al.** (2009). Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.* **151**: 496–505.
- Mickelson-Young, L., Endo, T.R., and Gill, B.S.** (1995). A cytogenetic ladder-map of the wheat homoeologous group-4 chromosomes. *Theor. Appl. Genet.* **90**: 1007–1011.
- Miftahudin, R.K., et al.** (2004). Analysis of expressed sequence tag loci on wheat chromosome group 4. *Genetics* **168**: 651–663.
- Moore, G., Devos, K.M., Wang, Z., and Gale, M.D.** (1995). Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.* **5**: 737–739.
- Murat, F., Xu, J.-H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., Messing, J., and Salse, J.** (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**: 1545–1557.

- Nei, M., and Gojobori, T. (1986). Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**: 418–426.
- Nelson, J.C., Sorrells, M.E., Van Deynze, A.E., Lu, Y.H., Atkinson, M., Bernard, M., Leroy, P., Faris, J.D., and Anderson, J.A. (1995). Molecular mapping of wheat: Major genes and rearrangements in homoeologous groups 4, 5, and 7. *Genetics* **141**: 721–731.
- Neumann, P., Pozárková, D., Vrána, J., Doležel, J., and Macas, J. (2002). Chromosome sorting and PCR-based physical mapping in pea (*Pisum sativum* L.). *Chromosome Res.* **10**: 63–71.
- Ozkan, H., Levy, A.A., and Feldman, M. (2001). Allopolyploidy-induced rapid genome evolution in the wheat (*Aegilops-Triticum*) group. *Plant Cell* **13**: 1735–1747.
- Paterson, A.H., et al. (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556.
- Potokina, E., Druka, A., Luo, Z., Wise, R., Waugh, R., and Kearsley, M. (2008). Gene expression quantitative trait locus analysis of 16 000 barley genes reveals a complex pattern of genome-wide transcriptional regulation. *Plant J.* **53**: 90–101.
- Qi, L., Friebe, B., and Gill, B.S. (2006). Complex genome rearrangements reveal evolutionary dynamics of pericentromeric regions in the Triticeae. *Genome* **49**: 1628–1639.
- Qi, L.L., et al. (2004). A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**: 701–712.
- Rosin, F.M., and Kramer, E.M. (2009). Old dogs, new tricks: Regulatory evolution in conserved genetic modules leads to novel morphologies in plants. *Dev. Biol.* **332**: 25–35.
- Salse, J., Abrouk, M., Bolot, S., Guilhot, N., Courcelle, E., Faraut, T., Waugh, R., Close, T.J., Messing, J., and Feuillet, C. (2009b). Reconstruction of monocotyledonous proto-chromosomes reveals faster evolution in plants than in animals. *Proc. Natl. Acad. Sci. USA* **106**: 14908–14913.
- Salse, J., Abrouk, M., Murat, F., Quraishi, U.M., and Feuillet, C. (2009a). Improved criteria and comparative genomics tool provide new insights into grass paleogenomics. *Brief. Bioinform.* **10**: 619–630.
- Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegou, B., Quraishi, U.M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C. (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**: 11–24.
- SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., and Bennetzen, J.L. (1998). The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **20**: 43–45.
- Sato, K., Nankaku, N., and Takeda, K. (2009a). A high-density transcript linkage map of barley derived from a single population. *Heredity* **103**: 110–117.
- Sato, K., Shin-I, T., Seki, M., Shinozaki, K., Yoshida, H., Takeda, K., Yamazaki, Y., Conte, M., and Kohara, Y. (2009b). Development of 5006 full-length cDNAs in barley: A tool for accessing cereal genomics resources. *DNA Res.* **16**: 81–89.
- Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jørgensen, J.-E., Weigel, D., and Andersen, S.U. (2009). SHOREmap: Simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6**: 550–551.
- Schulte, D., Close, T.J., Graner, A., Langridge, P., Matsumoto, T., Muehlbauer, G., Sato, K., Schulman, A.H., Waugh, R., Wise, R.P., and Stein, N. (2009). The international barley sequencing consortium —At the threshold of efficient access to the barley genome. *Plant Physiol.* **149**: 142–147.
- Shaked, H., Kashkush, K., Ozkan, H., Feldman, M., and Levy, A.A. (2001). Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridization and allopolyploidy in wheat. *Plant Cell* **13**: 1749–1759.
- Šimková, H., Svensson, J.T., Condamine, P., Hribová, E., Suchánková, P., Bhat, P.R., Bartoš, J., Safár, J., Close, T.J., and Doležel, J. (2008). Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* **9**: 294.
- Stein, N., Prasad, M., Scholz, U., Thiel, T., Zhang, H., Wolf, M., Kota, R., Varshney, R.K., Perovic, D., Grosse, I., and Graner, A. (2007). A 1,000-loci transcript map of the barley genome: New anchoring points for integrative grass genomics. *Theor. Appl. Genet.* **114**: 823–839.
- Steuernagel, B., et al. (2009). De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley. *BMC Genomics* **10**: 547.
- Suchánková, P., Kubaláková, M., Kovárová, P., Bartoš, J., Cíhalíková, J., Molnár-Láng, M., Endo, T.R., and Doležel, J. (2006). Dissection of the nuclear genome of barley by chromosome flow sorting. *Theor. Appl. Genet.* **113**: 651–659.
- Thiel, T., Graner, A., Waugh, R., Grosse, I., Close, T.J., and Stein, N. (2009). Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. *BMC Evol. Biol.* **9**: 209.
- Turner, A., Beales, J., Faure, S., Dunford, R.P., and Laurie, D.A. (2005). The pseudo-response regulator *Ppd-H1* provides adaptation to photoperiod in barley. *Science* **310**: 1031–1034.
- Vláčilová, K., Ohri, D., Vrána, J., Cíhalíková, J., Kubaláková, M., Kahl, G., and Doležel, J. (2002). Development of flow cytogenetics and physical genome mapping in chickpea (*Cicer arietinum* L.). *Chromosome Res.* **10**: 695–706.
- Waugh, R., Jannink, J.-L., Muehlbauer, G.J., and Ramsay, L. (2009). The emergence of whole genome association scans in barley. *Curr. Opin. Plant Biol.* **12**: 218–222.
- Wicker, T., Narechania, A., Sabot, F., Stein, J., Vu, G.T.H., Graner, A., Ware, D., and Stein, N. (2008). Low-pass shotgun sequencing of the barley genome facilitates rapid identification of genes, conserved non-coding sequences and novel repeats. *BMC Genomics* **9**: 518.
- Wicker, T., Schlagenhauf, E., Graner, A., Close, T.J., Keller, B., and Stein, N. (2006). 454 sequencing put to the test using the complex genome of barley. *BMC Genomics* **7**: 275.
- Wicker, T., Taudien, S., Houben, A., Keller, B., Graner, A., Platzer, M., and Stein, N. (2009). 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**: 712–722.
- Wu, J., Kurten, E.L., Monshausen, G., Hummel, G.M., Gilroy, S., and Baldwin, I.T. (2007). NaRALF, a peptide signal essential for the regulation of root hair tip apoplastic pH in *Nicotiana attenuata*, is required for root hair development and plant growth in native soils. *Plant J.* **52**: 877–890.
- Yan, L., Fu, D., Li, C., Blechl, A., Tranquilli, G., Bonafede, M., Sanchez, A., Valarik, M., Yasuda, S., and Dubcovsky, J. (2006). The wheat and barley vernalization gene VRN3 is an orthologue of FT. *Proc. Natl. Acad. Sci. USA* **103**: 19581–19586.
- Yang, Y., Dudoit, S., Luu, P., Lin, D., Peng, V., Ngai, J., and Speed, T. (2002). Normalization for cDNA microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**: e15.
- Yang, Z. (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**: 1586–1591.
- Zhang, G.Y., Wu, J., and Wang, X.W. (2010). Cloning and expression analysis of a pollen preferential rapid alkalization factor gene, *BoRALF1*, from broccoli flowers. *Mol. Biol. Rep.* **37**: 3273–3281.
- Zhou, F.S., Kurth, J.C., Wei, F.S., Elliott, C., Valé, G., Yahiaoui, N., Keller, B., Somerville, S., Wise, R., and Schulze-Lefert, P. (2001). Cell-autonomous expression of barley Mla1 confers race-specific resistance to the powdery mildew fungus via a Rar1-independent signaling pathway. *Plant Cell* **13**: 337–350.

Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences

Mihaela Maria Martis^a, Sonja Klemme^b, Ali Mohammad Banaei-Moghaddam^b, Frank R. Blattner^b, Jiří Macas^c, Thomas Schmutzer^b, Uwe Scholz^b, Heidrun Gundlach^a, Thomas Wicker^d, Hana Šimková^e, Petr Novák^c, Pavel Neumann^c, Marie Kubaláková^e, Eva Bauer^f, Grit Haseneyer^f, Jörg Fuchs^b, Jaroslav Doležel^e, Nils Stein^b, Klaus F. X. Mayer^a, and Andreas Houben^{b,1}

^aInstitute of Bioinformatics and Systems Biology/Munich Information Center for Protein Sequences, Helmholtz Center Munich, German Research Center for Environmental Health, 85764 Neuherberg, Germany; ^bLeibniz Institute of Plant Genetics and Crop Plant Research, 06466 Gatersleben, Germany; ^cBiology Centre, Academy of Sciences of the Czech Republic, Institute of Plant Molecular Biology, České Budějovice 37005, Czech Republic; ^dInstitute of Plant Biology, University of Zurich, 8008 Zurich, Switzerland; ^eCenter of the Region Haná for Biotechnological and Agricultural Research, Olomouc Research Center, Institute of Experimental Botany, Olomouc 77200, Czech Republic; and ^fDivision of Plant Breeding and Applied Genetics, Technical University of Munich, 85354 Freising, Germany

Edited by James A. Birchler, University of Missouri, Columbia, MO, and approved July 6, 2012 (received for review March 13, 2012)

Supernumerary B chromosomes are optional additions to the basic set of A chromosomes, and occur in all eukaryotic groups. They differ from the basic complement in morphology, pairing behavior, and inheritance and are not required for normal growth and development. The current view is that B chromosomes are parasitic elements comparable to selfish DNA, like transposons. In contrast to transposons, they are autonomously inherited independent of the host genome and have their own mechanisms of mitotic or meiotic drive. Although B chromosomes were first described a century ago, little is known about their origin and molecular makeup. The widely accepted view is that they are derived from fragments of A chromosomes and/or generated in response to interspecific hybridization. Through next-generation sequencing of sorted A and B chromosomes, we show that B chromosomes of rye are rich in gene-derived sequences, allowing us to trace their origin to fragments of A chromosomes, with the largest parts corresponding to rye chromosomes 3R and 7R. Compared with A chromosomes, B chromosomes were also found to accumulate large amounts of specific repeats and insertions of organellar DNA. The origin of rye B chromosomes occurred an estimated ~1.1–1.3 Mya, overlapping in time with the onset of the genus *Secale* (1.7 Mya). We propose a comprehensive model of B chromosome evolution, including its origin by recombination of several A chromosomes followed by capturing of additional A-derived and organellar sequences and amplification of B-specific repeats.

centromere | genome evolution | promiscuous DNA | non-Mendelian chromosome transmission

Supernumerary B chromosomes are not required for the normal growth and development of organisms and are assumed to represent a specific type of selfish genetic element. B chromosomes do not pair with any of the standard A chromosomes at meiosis, and have irregular modes of inheritance. Because they are dispensable for normal growth, B chromosomes have been considered non-functional, with no essential genes. As a result, B chromosomes follow their own species-specific evolutionary pathways. Despite their widespread occurrence in all eukaryotic groups, including insects (1), mammals (2), and plants (3), and their potential as chromosome-based vectors in biotechnology (4), little is known about the origin and molecular composition of these constituents of the genome.

Several scenarios have been proposed for the origin of B chromosomes. The most widely accepted view is that they are derived from the A chromosome complement. Some evidence also suggests that B chromosomes can be spontaneously generated in response to the new genomic conditions after interspecific hybridization. The involvement of sex chromosomes has also been argued for their origin in some species (reviewed in refs. 5–7). Despite the high number of species with B chromosomes, their de novo formation is

probably a rare event; the occurrence of similar B chromosome variants within related species suggests that they arose from a single origin.

One of the best-studied plant models for research into B chromosomes is rye (*Secale cereale*), with a genome comprising seven pairs of A chromosomes (1C ~7,917 Mbp) and containing between zero and eight B chromosomes, each with 1C ~580 Mbp. Rye B chromosomes appear to be monophyletic and very stable, being quite similar among rye taxa like *S. cereale* subsp. *segetale*, which is very closely related to *S. ancestrale* (8). This is rather unusual, given that B chromosomes are expected to have an elevated mutation rate compared with the A genome and thus should quickly diverge. At the DNA level, apart from the terminal region of the B chromosome long arm, overall the A and B chromosomes of rye are highly similar (9, 10). The molecular processes that gave rise to Bs during evolution remain unclear, and the characterization of sequences residing on them might shed light on their origin and evolution.

Our analysis provides insight into an enigmatic phenomenon of genome evolution in numerous groups of eukaryotes. We report that B chromosomes of rye are unexpectedly rich in gene-derived sequences, allowing us to trace their origin to parts of the A genome. In addition, compared with A chromosomes, B chromosomes accumulate large amounts of specific repeats and insertions of organellar DNA. We propose a model of the stepwise evolution of B chromosomes after segmental genome duplication followed by the capture of additional A-derived and organellar sequences and amplification of B-specific repeats.

Results

B Chromosomes Are Unexpectedly Rich in A-Derived Genic Sequences.

To identify the origin and evolution of the B chromosome, we performed a comparative sequence analysis of the A and B chromosomes of rye. First, we purified the B chromosome of an isogenic rye line by flow cytometry sorting (Fig. S1) and shotgun sequenced it at 0.9-fold sequence coverage using Roche 454 technology. As

Author contributions: F.R.B., J.M., U.S., N.S., K.F.X.M., and A.H. designed research; M.M.M., S.K., A.M.B.-M., F.R.B., T.S., H.G., T.W., H.Š., P. Novák, P. Neumann, M.K., E.B., G.H., J.F., and J.D. performed research; M.M.M., S.K., A.M.B.-M., F.R.B., T.S., H.G., T.W., H.Š., P. Novák, P. Neumann, M.K., E.B., G.H., J.F., J.D., and K.F.X.M. analyzed data; and M.M.M., S.K., A.M.B.-M., F.R.B., J.M., T.S., H.G., H.Š., P. Novák, P. Neumann, M.K., E.B., G.H., J.F., J.D., N.S., and A.H. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Database deposition: The sequences reported in this paper have been deposited in the European Nucleotide Archive database, <http://www.ebi.ac.uk/ena/> (accession no. ERP001061).

¹To whom correspondence should be addressed. E-mail: houben@ipk-gatersleben.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1204237109/-/DCSupplemental.

a reference, we used the sequence information from all A chromosomes (also purified by flow cytometry sorting) and the genomic DNA of plants both with (+B) and without (0B) B chromosomes (Table S1).

B chromosomes are generally considered nonfunctional, with no essential genes (5–7). Unexpectedly, we found many B sequences with a high homology to the genes of sequenced plant genomes (Fig. S2). Comparison of sequence reads from the rye B chromosome with the estimated size of 580 Mbp (BLASTX $\geq 70\%$ identity ≥ 30 amino acids) revealed a total of 4,189, 3,449, and 3,815 homologous nonredundant genes for *Brachypodium distachyon*, rice (*Oryza sativa*), and sorghum (*Sorghum bicolor*), respectively (Table S2). From the comparison of different individual reference datasets, a nonredundant gene count was extracted, comprising at least 4,946 putative B-located genic sequences. In comparison, the short arm of rye A chromosome 1R, with a size of 441 Mbp, is expected to contain $\sim 2,000$ genes (11). However, our analysis does not allow for conclusions regarding the completeness and functionality of the B-located genes.

We made use of the similarity between shared genic sequences of rye A and B chromosomes to determine the mutation frequency and relative age of the B-located sequences. To analyze the differences in SNP frequencies, which should reflect the presence or absence of selective pressure, we compared genic sequences from the As and Bs to rye RNAseq-based contigs (12) by BLASTN and identified SNPs in regions present in all three of the datasets. As expected, the genic sequences of rye A chromosomes revealed a lower SNP frequency (1/72 bp) than their B-located homologs (1/47 bp) compared with the rye RNAseq assemblies (Table S3). This difference is not related to a disparity in the effective population size between the chromosome sets, given that the SNP frequencies of mobile elements were one SNP per 25 bp in the A chromosomes and one SNP per 26 bp in the B chromosomes. Thus, the selection pressure is lower for B-located genes than for A-located genes.

We used sequence alignments of the A and B gene sequences and their homologs in *Brachypodium* and barley full-length cDNAs in Bayesian phylogenetic analyses to determine the age of origin of the B chromosome. The inferred age of 1.1–1.3 million y (My) of rye B chromosomes (Fig. S3) might be overestimated owing to the relaxed selective pressure on B-located genes. Nevertheless, it coincides with the estimated age of 1.7 My for the genus *Secale* and 0.8 My for the *S. strictum/S. vavilovii/S. cereale* taxon group, based on a dated rDNA phylogeny of Triticeae (Fig. S4). These ages indicate that rye B chromosomes originated only within the genus *Secale* and are in accord with the present-day occurrence of B chromosomes in *S. cereale* alone.

The identification of B-sequence reads with similarity to conserved coding sequences and the close syntenic relationship among grass genomes allowed us to trace the putative chromosomal origin of rye B sequences. Using a stringent filter criterion of ≥ 30 amino acids/100 bp similarity, we analyzed the rye B reads against the virtual gene map of barley (13) and the assembled genomes of *Brachypodium*, rice, and sorghum to depict the positional information on the respective chromosomes. Rye B chromosomes apparently contain several prominent blocks of conserved genes corresponding to barley chromosomal regions 2H, 3H, 4H, and 5H, along with thousands of short genic sequences scattered all over the A chromosomes (Fig. 1A). In contrast, reads from the short arm of rye A chromosome 1R (1RS) corresponded mainly to the syntenic barley chromosome 1H (Fig. S5). These results indicate that the randomly scattered pattern observed for rye B sequences is exclusive to B chromosomes and is not shared by A chromosomes. A comparison of the reads with the sequences of *Brachypodium* and sorghum confirmed the genome-wide scattered distribution of the rye B reads.

B Chromosomes Accumulate Large Amounts of Organellar Sequences. The rye A and B chromosomes were further compared with respect to the content and frequency of individual classes of

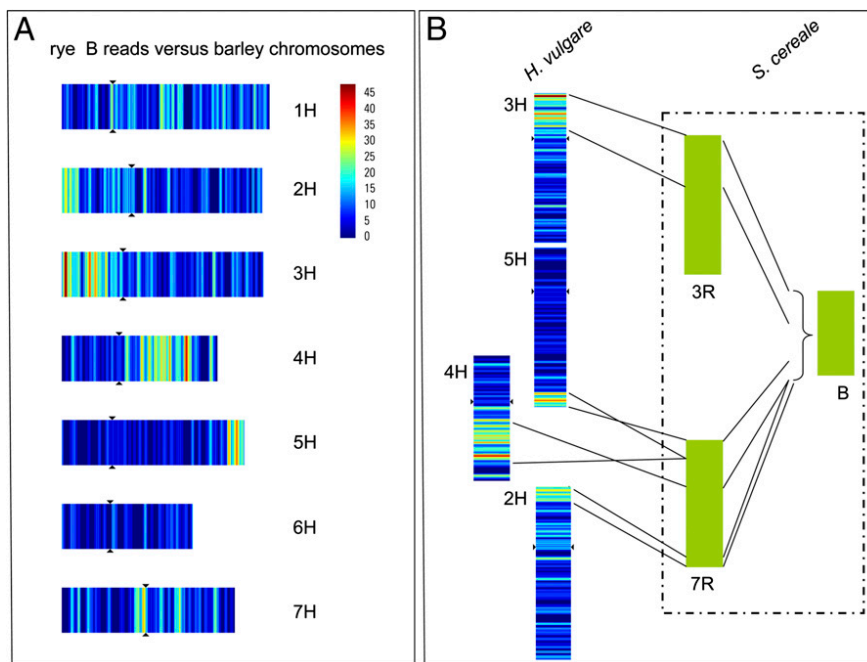


Fig. 1. Multichromosomal origin of the rye B chromosome. (A) Rye B sequence reads mapped onto the barley genome. The heatmap depicts the detected homologous (syntenic) regions in the barley genome. Sequence reads were anchored on barley chromosomes 1H–7H using BLASTN and the best detectable match. Individual chromosomes are numbered. Multiple regions exhibit conserved genes with respect to barley chromosomes 2H, 3H, 4H, and 5H (implied syntenic regions) and multiple small regions on the remaining chromosomes. (B) Syntenic relationship between the A chromosomes of barley and rye. Chromosome 7R corresponds to regions from 2H, 4H, 5H, and 7H (15). Thus, the B chromosome of rye shows extended similarity to regions of chromosomes 3R and 7R.

repeats. Repeat identification using similarity-based clustering of sequence reads (14) revealed that almost 90% of the rye genome is composed of repetitive DNA, and 70% of the genome is represented by fewer than 60 different repeat families. Although the B chromosomes contained a similar proportion of repeats as the A chromosomes, the two differed significantly in composition owing to an additional massive accumulation of B-specific satellite repeats (Fig. S6 and Table S4). The B-specific satellite repeats were characterized by exceptionally long monomers (0.9–4.0 kb), and their partial similarity to other types of repeats suggests chimeric origins. In addition to satellite repeats, an accumulation of sequences corresponding to the *Bianka* family of Ty1/*copia* elements was also observed in the B chromosomes.

Furthermore, B chromosomes accumulated significant amounts of plastid (NUPT)- and of mitochondrion (NUMT)-derived sequences. All parts of the plastid and mitochondrion genomes were transferred to the B chromosomes, indicating that all sequences are transferable. The higher number of organelle-derived DNA inserts in B chromosomes than in A chromosomes (Fig. S7) and the increased mutation frequency of B-located organellar DNA suggest a reduced selection against organellar DNA in supernumerary chromosomes. We also observed that along with large amounts of mitochondrion-derived DNA, B-enriched high-copy repeats are integrated in the centromeric region (Fig. 2). Therefore, the centromere might facilitate the evolution of the B chromosomes by accumulation and shuffling of sequences. Whether the distinct centromere composition of the B chromosomes plays a role in the B-specific drive mechanism, resulting in non-Mendelian chromosomal segregation behavior, remains to be tested.

Discussion

We have described a unique comprehensive model of B chromosome evolution based on comparative sequence analysis of the A and B chromosomes of rye. Considering the similar age of the genus *Secale* and the age of its B chromosomes, it is tempting to

speculate that B chromosomes originated as a by-product of chromosome rearrangement events. This hypothesis is supported by the notion that the rye genome underwent a series of rearrangements after its split from the wheat and barley lineages and as such is an exception to otherwise pronounced genome colinearity in Triticeae (15). Thus, chromosomes 3R of rye and 3H of barley are mainly conserved and syntenic to each other, whereas rye 7R shares conserved synteny with barley chromosomal regions 2H, 4H, and 5H (Fig. S8). Based on the comparison of the rye B-specific sequence reads to the linear genome model of barley (13), we conclude that the rye B chromosomes originated primarily from the rye chromosomal regions 3RS and 7R after multiple chromosomal rearrangements (Fig. 1B). A multichromosomal origin of B-chromosome sequences is further supported by the many short sequences that are similar to other regions of the A chromosomes. A comparable amalgamation of diverse A-derived sequences has been previously postulated for the B chromosomes of maize (16) and *Brachycome dichromosomatica* (17). The intron-containing gene reads found among the B-sequence reads, corresponding to regions outside of the 3RS and 7R regions, might represent insertions into the B chromosomes that occurred during double-strand break repair (18) or results from hitchhiking genomic fragments with transposable elements, as demonstrated for noncollinear genes of Triticeae (19).

The most unexpected result of our analysis is the discovery that B chromosomes are rich in gene fragments that represent copies of A chromosome genes. Although our analysis does not allow us to draw any conclusions regarding the completeness and functionality of the B-located genes, preliminary analyses indicate that B-located sequences are transcribed only weakly (20). Considering the coexistence of sequence-identical A- and B-derived transcripts, it is likely that “dosage compensation” occurs in rye, with an equal expression regardless of the copy number of the respective gene. An efficient dosage compensation mechanism might explain the weak phenotype caused by B chromosomes.

What mechanism could account for the accumulation of organellar DNA in B chromosomes of rye? Transfer of organellar DNA to the nucleus is very frequent (21), but most of the “promiscuous” DNA is also rapidly lost again via a counterbalancing removal process (22). If this expulsion mechanism is impaired in B chromosomes, then the high turnover rates that prevent such sequences on the A chromosomes from accumulating and degrading would be absent and allow for sequence decay. Thus, the dynamic equilibrium between frequent integration and rapid elimination of organellar DNA could be imbalanced for B chromosomes. We also observed that the large amounts of mitochondrion-derived DNA integrated preferentially in the B pericentromeric region. Pericentromeric regions generally contain few functional genes, and this low gene density may facilitate the repeated integration of the organelle-derived DNA (23). Alternatively, consistent with the rapid evolution of centromeres (24) after sequence integration, subsequent amplification of these sequences might have occurred within this region. Future analyses of other B-bearing species are needed to address the question whether organelle-to-nucleus DNA transfer is an important mechanism that drives the evolution of B chromosomes.

Based on our findings, we propose a multistep model for the origin of a selfish chromosome (Fig. 3). Initially, a proto-B chromosome was formed by segmental or whole-genome duplication, subsequent chromosome translocations, unbalanced segregation of a small translocation chromosome, and subsequent sequence insertions. The recombination with donor A chromosomes became restricted, likely owing to multiple rearrangements involving different A chromosomes, which no longer allowed extended pairing with the originally homologous A regions. This restriction of recombination can be considered the starting point for the independent evolution of B chromosomes. The presence of fast-evolving repetitive sequences, along with reduced selective

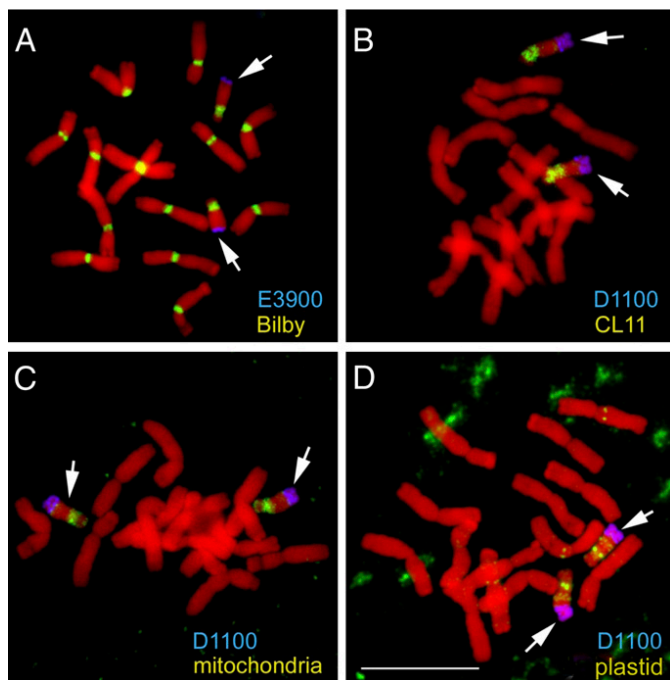


Fig. 2. FISH of rye mitotic metaphase chromosomes with the centromeric retrotransposons *Bilby* (A), the B-specific pericentromeric Ty1/*copia* repeat CL11 (B), mitochondrial DNA (C), and plastid DNA (D). B chromosome-specific satellite repeats E3900 and D1100 were used for identifications of the Bs. The Bs are indicated by arrows. (Scale bar: 10 μ m.)

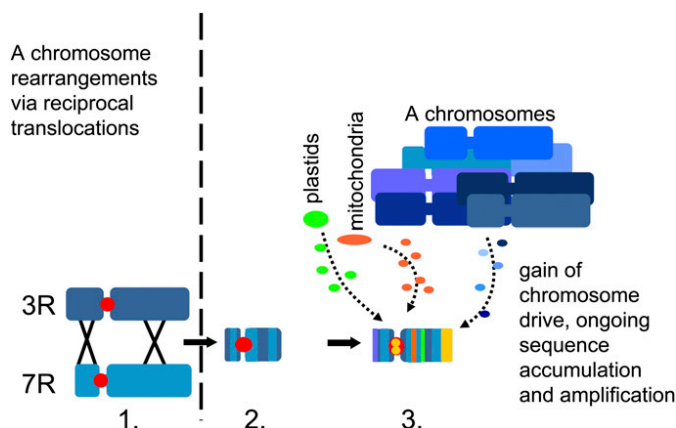


Fig. 3. Model of the stepwise evolution of the rye B chromosome after segmental genome duplication. (1) Reciprocal translocation of duplicated fragments of the 3R and 7R chromosomes and unbalanced segregation of a small translocation chromosome results in (2) decay of meiotic A–B pairing and the formation of a proto-B. (3) The accumulation of organellar and A chromosome-derived DNA fragments, amplification of B-specific repeats, erosion and inactivation of A-derived genes (Muller's ratchet), and gain of chromosome drive resulted in the B chromosome.

pressure on gene integrity, could predispose a nascent B chromosome to undergo further rapid structural modifications required to establish a drive mechanism. Because an increased gene dosage may affect gene expression, the expression of paralogues on B chromosomes might have been reprogrammed (potentially through epigenetic mechanisms) early during the evolution of the B chromosomes. Thus, proto-B genes might have first been suppressed by silencing mechanisms and then degenerated owing to mutations and the insertion of sequences derived from other A-chromosomal regions and organellar genomes, except for those coding and/or noncoding sequences providing drive and an advantage for the maintenance of B chromosomes. Our model predicts that B chromosomes occur primarily in taxa with elevated levels of chromosomal rearrangements and phylogenetic groups with unstable chromosome numbers.

- Wilson EB (1907) The supernumerary chromosomes of *Hemiptera*. *Science* 26:870.
- Hayman DL, Martin PG (1965) Supernumerary chromosomes in the marsupial *Chionobates volans* (Kerr). *Aust J Biol Sci* 18:1081–1082.
- Longley AE (1927) Supernumerary chromosomes in *Zea mays*. *J Agric Res* 35:769–784.
- Yu W, Lamb JC, Han F, Birchler JA (2006) Telomere-mediated chromosomal truncation in maize. *Proc Natl Acad Sci USA* 103:17331–17336.
- Camacho JPM, Sharbel TF, Beukeboom LW (2000) B-chromosome evolution. *Philos Trans R Soc Lond B Biol Sci* 355:163–178.
- Jones N, Houben A (2003) B chromosomes in plants: Escapees from the A chromosome genome? *Trends Plant Sci* 8:417–423.
- Burt A, Trivers R (2006) *Genes in Conflict: The Biology of Selfish Genetic Elements* (Belknap Press of Harvard Univ Press, Cambridge, MA), p 602.
- Niwa K, Sakamoto S (1995) Origin of B chromosomes in cultivated rye. *Genome* 38:307–312.
- Timmis JN, Ingle J, Sinclair J, Jones RN (1975) Genomic quality of rye B chromosomes. *J Exp Bot* 26:367–378.
- Tsujimoto H, Niwa K (1992) DNA structure of the B chromosome of rye revealed by in situ hybridization using repetitive sequences. *Jpn J Genet* 67:233–241.
- Kubaláková M, et al. (2003) Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. *Genome* 46:893–905.
- Haseneyer G, et al. (2011) From RNA-seq to large-scale genotyping: Genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol* 11:131.
- Mayer KF, et al. (2011) Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* 23:1249–1263.
- Novák P, Neumann P, Macas J (2010) Graph-based clustering and characterization of repetitive sequences in next-generation sequencing data. *BMC Bioinformatics* 11:378.
- Devos KM, et al. (1993) Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor Appl Genet* 85:673–680.
- Peng SF, Lin YP, Lin BY (2005) Characterization of AFLP sequences from regions of maize B chromosome defined by 12 B-10L translocations. *Genetics* 169:375–388.

Materials and Methods

Purification of Mitotic Chromosomes and 454 Sequencing. A and B chromosomes of rye (*S. cereale*) inbred line 7415 (25) were isolated by flow cytometry sorting and shotgun sequenced by Roche 454 (11, 13).

Analysis of Repetitive DNA and Organellar DNA Insertions. The content of the repetitive DNA per sequence read was identified by Vmatch (<http://www.vmatch.de>) against the MIPS-REdat Poaceae v8.6.1 repeat library. The clustering analysis of sequence reads was performed as described previously (14). The A and B sequence reads were compared (BLASTN) against the plastid and mitochondrial genomes of wheat (AB042240 and AP008982).

Identification of Gene Reads and Comparative Genomics. Gene numbers were estimated by BLAST comparisons with the repeat-filtered reads against the proteins/coding sequences of *B. distachyon*, rice (*O. sativa*), and sorghum (*S. bicolor*) and against EST collections. Rye 1RS (EMBL-EBI European Bioinformatics Institute, <http://www.ebi.ac.uk>, NCBI accession no. SRX019678) and B chromosome datasets were compared with reference genomes (BLASTX) as described previously (13).

Detection of SNPs and Dating of Rye B Chromosome Origin. Genic 454 shotgun reads from A and B chromosomes were mapped against matching sequences from the rye transcriptome dataset using BWA (26). An SNP-based comparison of A and B chromosomes was performed to date the age of the rye B. The regions that contained high-quality SNPs in A and B were mapped onto the corresponding barley full-length cDNAs (26, 27) and *Brachypodium* reference genome version 1.2 using BLASTN. The datasets were phylogenetically analyzed with Bayesian inference in MrBayes 3.1.2 (28) and dated in BEAST 1.6.1 (29). More detailed descriptions of methods are provided in *SI Materials and Methods*.

ACKNOWLEDGMENTS. We thank I. Schubert, R. N. Jones, M. Puertas, J. Timmis, D. Weigel, J. Birchler, and B. Steuernagel for fruitful discussions; S. König for excellent technical support of the 454 sequencing of the rye A and B chromosomes; K. Burg for sequence information for 1RS; and J. Číhalíková, R. Šperková, and Z. Dubská for their help with chromosome sorting. This work was supported by the German Research Foundation Grant HO 1779/10-1/14-1; German Federal Ministry of Education and Research Grant FKZ 0315063B (Tritex and Gabi Rye Express Projects 0315954C and 0315063C); Czech Science Foundation Grant P501/12/G090; Ministry of Education, Youth and Sports of the Czech Republic Grant OC10037; European Regional Development Fund Operational Programme Research and Development for Innovations Grant ED0007/01/01; and Academy of Sciences of the Czech Republic Grant AVOZ50510513.

- Houben A, Verlin D, Leach CR, Timmis JN (2001) The genomic complexity of micro B chromosomes of *Brachycome dichromosomatica*. *Chromosoma* 110:451–459.
- Salomon S, Puchta H (1998) Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *EMBO J* 17:6086–6095.
- Wicker T, et al. (2011) Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* 23:1706–1718.
- Carchilan M, Kumke K, Mikolajewski S, Houben A (2009) Rye B chromosomes are weakly transcribed and might alter the transcriptional activity of A chromosome sequences. *Chromosoma* 118:607–616.
- Timmis JN, Ayliffe MA, Huang CY, Martin W (2004) Endosymbiotic gene transfer: Organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5:123–135.
- Sheppard AE, Timmis JN (2009) Instability of plastid DNA in the nuclear genome. *PLoS Genet* 5:e1000323.
- Matsuo M, Ito Y, Yamauchi R, Obokata J (2005) The rice nuclear genome continuously integrates, shuffles, and eliminates the chloroplast genome to cause chloroplast-nuclear DNA flux. *Plant Cell* 17:665–675.
- Hall AE, Keith KC, Hall SE, Copenhaver GP, Preuss D (2004) The rapidly evolving field of plant centromeres. *Curr Opin Plant Biol* 7:108–114.
- Jimenez MM, Romera F, Puertas MJ, Jones RN (1994) B-chromosomes in inbred lines of rye (*Secale cereale* L.). 1: Vigor and fertility. *Genetica* 92:149–154.
- Matsumoto T, et al. (2011) Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol* 156:20–28.
- Sato K, et al. (2009) Development of 5006 full-length cDNAs in barley: A tool for accessing cereal genomics resources. *DNA Res* 16:81–89.
- Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19:1572–1574.
- Drummond AJ, Rambaut A (2007) BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* 7:214.

LARGE-SCALE BIOLOGY ARTICLE

Reticulate Evolution of the Rye Genome ^{WJOPEN}

Mihaela M. Martis,^{a,1} Ruonan Zhou,^{b,1} Grit Haseneyer,^c Thomas Schmutzer,^b Jan Vrána,^d Marie Kubaláková,^d Susanne König,^b Karl G. Kugler,^a Uwe Scholz,^b Bernd Hackauf,^e Viktor Korzun,^f Chris-Carolin Schön,^c Jaroslav Doležel,^d Eva Bauer,^c Klaus F.X. Mayer,^a and Nils Stein^{b,2}

^aHelmholtz Center Munich, German Research Centre for Environmental Health, Munich Information Center for Protein Sequences/IBIS, Institute of Bioinformatics and Systems Biology, 85764 Neuherberg, Germany

^bLeibniz Institute of Plant Genetics and Crop Plant Research, 06466 Seeland (OT) Gatersleben, Germany

^cTechnische Universität München, Centre of Life and Food Sciences Weihenstephan, Plant Breeding, 85354 Freising, Germany

^dCentre of the Region Haná for Biotechnological and Agricultural Research, Institute of Experimental Botany, CZ-783 71 Olomouc, Czech Republic

^eJulius Kühn-Institut, Institute for Breeding Research on Agricultural Crops, 18190 Sanitz, Germany

^fKWS LOCHOW, 29296 Bergen, Germany

ORCID ID: 0000-0003-3011-8731 (N.S.).

Rye (*Secale cereale*) is closely related to wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*). Due to its large genome (~8 Gb) and its regional importance, genome analysis of rye has lagged behind other cereals. Here, we established a virtual linear gene order model (genome zipper) comprising 22,426 or 72% of the detected set of 31,008 rye genes. This was achieved by high-throughput transcript mapping, chromosome survey sequencing, and integration of conserved synteny information of three sequenced model grass genomes (*Brachypodium distachyon*, rice [*Oryza sativa*], and sorghum [*Sorghum bicolor*]). This enabled a genome-wide high-density comparative analysis of rye/barley/model grass genome synteny. Seventeen conserved syntenic linkage blocks making up the rye and barley genomes were defined in comparison to model grass genomes. Six major translocations shaped the modern rye genome in comparison to a putative Triticeae ancestral genome. Strikingly dissimilar conserved syntenic gene content, gene sequence diversity signatures, and phylogenetic networks were found for individual rye syntenic blocks. This indicates that introgressive hybridizations (diploid or polyploidy hybrid speciation) and/or a series of whole-genome or chromosome duplications played a role in rye speciation and genome evolution.

INTRODUCTION

Rye (*Secale cereale*) is a member of the Triticeae tribe of the Pooideae subfamily of grasses. It is closely related to wheat (*Triticum aestivum*) and barley (*Hordeum vulgare*) and provides a main cereal for food and feed in Eastern and Northern Europe. Rye, in contrast with wheat and barley, is allogamous, and reproduction is controlled by a bifactorial self-incompatibility system promoting outcrossing (Lundqvist, 1956). A combination of male sterility inducing cytoplasm and nuclear-encoded fertility-restorer genes forms the basis of efficient hybrid breeding in rye for improved exploitation of heterosis (Geiger and Miedaner, 2009). Elevated abiotic stress tolerance to frost, drought, and marginal soil fertility make rye a perfect model for functional

analyses and consequently improvement of cereal crops like wheat and barley, which are less tolerant to abiotic stress.

Rye has a large (1C = 8.1 Gb; Doležel et al., 1998) diploid genome (2n = 2x = 14), nearly 50% bigger than the barley genome. It is unknown whether this results from higher amounts of repetitive DNA only or if rye also contains more genes than other diploid Triticeae species. Similar to wheat and barley, the center of origin of genus *Secale* is in the Near East. Rye was domesticated during the Neolithic Era (7000 years ago) in Anatolia and later in Europe, where it first spread as a weed in wheat and barley fields (Sencer and Hawkes, 1980; Willcox, 2005). Rye and wheat diverged seven million years ago, and both lineages and the barley lineage diverged from a common Triticeae ancestor around 11 million years ago (Huang et al., 2002).

Despite extensive synteny to barley (H genome) and wheat (A, B, and D genomes), the rye genome (R) has undergone a series of rearrangements, as revealed by comparative restriction fragment length polymorphism (RFLP) mapping (Devos et al., 1993). Colinearity to wheat was disturbed by a series of translocations involving all chromosomes but 1R. It was postulated that a translocation involving the long arms of linkage groups 4 and 5 (4L/5L) occurred before the split of the wheat and rye lineages, since it is present in various Triticeae species and in the A genome of wheat (Moore et al., 1995; Mayer et al., 2011). Subsequent reorganization events involving several other chromosome arms

¹ These authors contributed equally to this work.

² Address correspondence to stein@ipk-gatersleben.de.

The authors responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) are: and Klaus F.X. Mayer (k.mayer@helmholtz-muenchen.de) and Nils Stein (stein@ipk-gatersleben.de).

^{WJOPEN} Online version contains Web-only data.

^{WJOPEN} Articles can be viewed online without a subscription. www.plantcell.org/cgi/doi/10.1105/tpc.113.114553

were proposed (Devos et al., 1993). Comprehensive genome-wide analysis of the level of conserved synteny and extension of rearrangements between rye and other Triticeae genomes has so far been hampered by lack of genomic resources in rye.

High-density gene-based marker maps are important prerequisites for studying genome organization and evolution. Such maps in barley (Stein et al., 2007; Close et al., 2009; Sato et al., 2009) and wheat (Qi et al., 2004) allowed detailed comparisons to sequenced model grass genomes like rice (*Oryza sativa*), *Brachypodium distachyon*, and sorghum (*Sorghum bicolor*) (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; International Brachypodium Initiative, 2010). A dense gene-based genetic map of barley together with conserved synteny information of the above mentioned three model grass genomes provided the framework to integrate a linear gene order model comprising more than 21,000 barley genes. The gene content information of barley was obtained by survey sequencing of amplified DNA from individually sorted chromosomes (Mayer et al., 2009; Mayer et al., 2011). Thus genome size, which hampered systematic sequencing of Triticeae genomes for long time, could be turned into an advantage in Triticeae genome analysis since chromosomes can be sorted and enriched from different Triticeae species including rye (Kubaláková et al., 2003; Doležel et al., 2012).

For rye, existing genetic maps comprised limited numbers of gene-based markers (Gustafson et al., 2009; Hackauf et al., 2009) or were composed of anonymous genomic Diversity Arrays Technology markers (Milczarski et al., 2011). Recently, a large data set of gene-based single nucleotide polymorphisms (SNPs) could be data-mined from RNA sequencing data of rye, providing the basis for developing a high-throughput SNP genotyping assay comprising 5234 markers (Haseneyer et al., 2011). In this study, this SNP assay was employed to build a high-density transcript map of rye. Together with chromosomal survey sequences (CSSs) generated from flow-sorted and amplified rye chromosomes, a high-density linear gene-order map could be established. This provided the basis for in-depth comparative genetic analysis between rye and other grass genomes, leading us to propose a revised model of rye genome evolution. Global sequence conservation and synteny and phylogenetic network analysis revealed a heterogeneous composition of the rye genome, indicating its reticulate evolution (evolutionary relationships do not fit a simple bifurcate tree but instead fit a network structure), which can be linked to a series of translocations that shaped the rye genome. We postulate that this was the result of introgressive hybridization and/or allopolyploidization events. The

outbreeding lifestyle of rye might have facilitated interspecies introgressive hybridization, thus providing an important prerequisite for the formation of the modern rye genome.

RESULTS

A High-Density Transcript Map of Rye

A high-density gene-based marker map of rye was developed by genotyping 495 recombinant inbred lines (RILs) from four mapping populations with a previously published Rye5K Infinium Bead Chip (Haseneyer et al., 2011) comprising 5234 SNP markers (Table 1). In addition, 271 Expressed Sequence Tag (EST)-SSR (for simple sequence repeat) markers were genotyped in two of the populations. Between 782 and 2158 SNP and SSR markers were mapped in the four individual mapping populations (Table 1). An integrated high-density genetic map comprising 3543 gene-based markers and 45 anchor markers (providing links to previous work in rye; Hackauf et al., 2012) was established, encompassing a cumulative map length of 1947 centimorgans (Figure 1; see Supplemental Figure 1 online).

Composition of Rye Chromosomes Revealed by Survey Sequencing

Individual rye chromosomes were purified and used as template for CSS using Roche/454 technology. We obtained between 1.02 (chromosome 1R) and 1.43 (4R) Gb of sequence per chromosome. In total, 8.25 Gb provided sequence coverage between 0.93- and 1.17-fold (average 1.04-fold) for each individual chromosome fraction (Table 2). The expected base pair coverage was calculated to range between 60.5 and 68.9% (average 64.6%; Table 2). The estimated values were tested by comparing the CSS data sets against the available genetically anchored sequence markers. An average marker detection rate (sensitivity) of 78.7% was observed, and for all individual chromosomes, the theoretically expected Lander-Waterman values were significantly exceeded. The average specificity of 92.6% (Table 2) correlated well with cytological estimates of the average individual chromosome fraction purity of 93.5% obtained by fluorescence in situ hybridization on specimens prepared from sorted chromosome fractions.

To identify the fraction of CSS reads containing gene and/or exon sequence, we masked all repetitive DNA sequences. About 74% of the CSS sequences consisted of repetitive DNA elements (see Supplemental Table 1 online). The remaining 2.2 Gb

Table 1. Molecular Marker Statistics for Transcript Mapping in Rye

Mapping Population ^a	EST-SNP	EST-SSR	Anchor Markers	No. of Mapped Markers	No. of Mapped Genes	Map Length (cM) ^b
Lo7xLo225	1952	206	–	2158	1825	1428
P87xP105	1813	–	–	1813	1504	1347
Lo90xLo115	717	65	–	782	677	1084
L2039-NxDH	1200	–	45	1245	1038	1369
Consensus	3272	271	45	3588	2886	1947

^aMaps generated with JoinMap v4.0, except P87xP105, which has been calculated with MSTMap.

^bcM, centimorgans. –, not available.

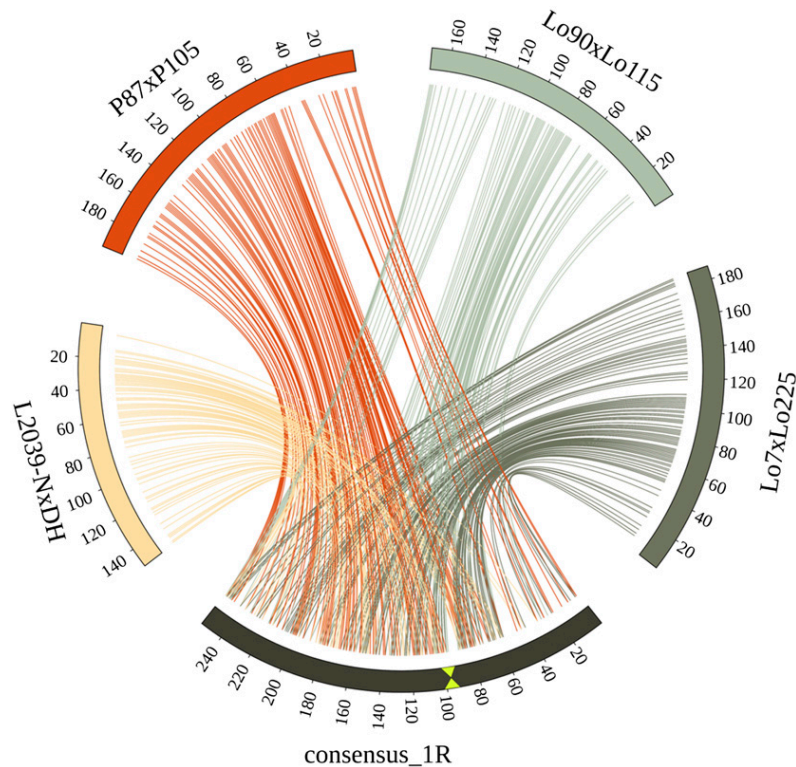


Figure 1. Rye Consensus Transcript Map.

Comparison of the integrated genetic map of chromosome 1R with the 1R maps of four individual mapping populations (Lo7xLo225, P87xP105, Lo90xLo115, and L2039-NxDH). Colored lines connect markers between the integrated map and each individual genetic linkage map. Complete collinearity could be observed between all individual maps and the integrated consensus. Centromere position in the consensus map is indicated by green triangles.

of sequence was distributed among the individual rye chromosomes resulting in a range between 275 Mb assigned to 7R and 437 Mb assigned to 4R. This repeat-masked CSS fraction was compared with a recently published set of barley genes (International Barley Genome Sequencing Consortium, 2012) and full gene sets of the sequenced genomes of rice, *B. distachyon*, and sorghum (International Rice Genome Sequencing Project, 2005; Paterson et al., 2009; International Brachypodium Initiative, 2010). Overall, sequence similarity was obtained for a non-redundant set of 31,008 genes. On the basis of the previously determined sensitivity of the sequence data sets, more than 39,400 genes thus can be estimated for the rye genome.

Virtual Linear Order of 22,426 Rye Genes (Genome Zipper)

Previously, we introduced the concept of developing virtual linear gene order maps (genome zippers) by integrating CSS data with dense gene-based marker maps and conserved synteny information from sequenced model grass genomes (i.e., *B. distachyon*, rice, and sorghum) (Mayer et al., 2009, 2011). We followed this approach for the rye CSS data. In the first step, a comparison of genes constituting the transcript map of rye established the putatively orthologous (conserved syntenic) regions of the model grass genomes. Subsequently, all coding

sequences from CSS data were compared against genes from these reference genomes. Based on genes located in corresponding syntenic blocks of the respective model grass genomes and identified with rye CSS data, it was postulated that the putatively orthologous genes are present in a conserved order in rye as well. Hence, the high-density transcript map of rye provided the scaffold to position and orient blocks of conserved syntenic genes between rye and the model grass genomes. A total of 10,833 barley cDNAs, 20,370 nonredundant rye ESTs, and between 11,869 and 14,086 genes from reference genomes (see above) were unambiguously associated with rye CSS sequences (Table 3). Between 2693 (6R) and 3595 (2R) genes were assigned in linear order along individual rye chromosomes (Table 3; see Supplemental Data Sets 1 to 7 online). Overall, 22,426 rye genes were positioned along the genome. Thus, we were able to position 72% of all detected rye genes (22,426/31,008).

Conserved Synteny between the Genomes of Rye and Barley

The close evolutionary relationship between rye and barley is reflected in extensively conserved synteny. On the basis of the above presented linear gene-order map of rye, structural

Table 2. Sequence and Coverage Statistics from CSSs of Individual Rye Chromosomes

Chromosome	Size (Mb) ^a	Sequences (Mb)	Coverage (x-Fold)	Expectation ^b	Observed Marker Detection Rate (Sensitivity)	Anchored Reads (Specificity)
1R	1005	1023	1.02	63.9	75.4	84.7
2R	1315	1253	0.95	61.3	80.2	95.7
3R	1047	1226	1.17	68.9	77.4	93.0
4R	1242	1435	1.16	68.6	80.7	93.4
5R	1119	1229	1.10	66.7	80.9	93.9
6R	1134	1060	0.93	60.5	76.4	94.3
7R	1055	1027	0.97	62.1	79.9	93.1
Total	(Σ) 7917	(Σ) 8253	(\bar{O}) 1.04	(\bar{O}) 64.6	(\bar{O}) 78.7	(\bar{O}) 92.6

^aCalculated based on 2C DNA amount = 16.19 pg (Doležel et al., 1998), relative chromosome lengths according to Schlegel et al. (1987), and 1 pg = 0.978 Mb (Doležel et al., 2003).

^bExpectation was calculated using the Lander Waterman expectation (Lander and Waterman, 1988).

differences, translocations, and the overall extent of conserved synteny could now be addressed at unprecedented resolution between rye and barley or the other reference grass genomes, respectively. Comparisons of the dense genetic rye map provided in this study and the physical/genetic barley genome assembly (International Barley Genome Sequencing Consortium, 2012) revealed numerous rearrangements in rye chromosomes (Figure 2; see Supplemental Figure 2 online). Only rye chromosome 1R exhibited collinearity over its entire length to a single barley chromosome (1H). All other rye chromosomes were composed of a mosaic pattern with two to four conserved syntenic segments of individual barley chromosomes (Figure 2; see Supplemental Figure 2 online). The 2R markers and 454 sequences of the genome zipper identified a small part corresponding to barley chromosome 7HL and almost the entire chromosome 2H. The 3R marker corresponded to almost the entire chromosome 3H and a region on 6HL, while 4R-tagged regions on 4H and segments from the short arms of 6H and 7H. Chromosome 5R tagged regions on 5H and 4HL. Chromosome 6R is homoeologous with most, but not all, of chromosome 6H and with the long arms of 3H and 7H. Chromosome 7R is composed of segments with homoeology to parts of 4HL, 5HL, and 7HL as well as to parts of 2HS and 7HS. All seven genetic centromeres in rye and barley (Figure 2) are conserved at syntenic positions and were not involved in translocations in rye. They thus remained conserved since the divergence of a common ancestor. Overall, we identified 17 conserved syntenic

segments between rye and barley that make up both genomes and allow us to propose a revised model of rye genome evolution (Figure 3). This model describes a series of six translocation events that account for the major pattern of rearrangements between rye and barley.

Conserved Synteny to Model Grass Genomes Is Nonuniform between Rye and Barley

Based on the extent of conserved synteny between rye and barley, we compared the global pattern of conserved synteny to sequenced model grass genomes. Overall, rye and barley contain very similar numbers of conserved syntenic genes when compared with *B. distachyon*, rice, and sorghum (see Supplemental Table 2 and Supplemental Figures 3 and 4 online; Figure 2). Comparing the rye (this study) and barley (Mayer et al., 2011) genome zippers, which are established by integrating synteny information with regard to the same three model grass genomes, both species share 64 to 66% (14,408) of the 22,426 and 21,766 respective genome zipper loci. Given the large number of rearrangements between the rye and barley genomes, we addressed the question whether all conserved syntenic blocks between both genomes contain proportional numbers of conserved syntenic genes in comparison to the three model grass genomes. We surveyed all 17 conserved syntenic regions between rye and barley individually. In most cases, barley and rye segments carried similar or equal numbers of conserved syntenic

Table 3. Genome Zipper Statistics: Genes, ESTs, and Associated 454 Reads

Data Sets	1R	2R	3R	4R	5R	6R	7R	Σ
No. of SNP markers	390	469	381	394	486	398	422	2,940
No. of markers with orthologous gene in reference genome(s)	224	270	223	215	276	199	236	1,643
No. of barley fl-cDNAs	1,386	1,663	1,567	1,437	1,697	1,370	1,713	10,833
No. of nonredundant sequence reads	23,720	29,907	24,948	36,818	33,671	21,436	24,304	194,804
No. of matched rye ESTs	2,489	3,121	2,849	2,892	3,382	2,877	2,760	20,370
No. of <i>B. distachyon</i> genes	1,761	2,291	2,146	1,960	2,391	1,750	1,787	14,086
No. of rice genes	1,469	2,060	1,825	1,510	1,767	1,444	1,794	11,869
No. of sorghum genes	1,538	1,818	2,015	1,644	2,050	1,439	1,740	12,244
No. of nonredundant anchored gene loci in genome zipper	2,806	3,595	3,201	3,299	3,751	2,693	3,081	22,426

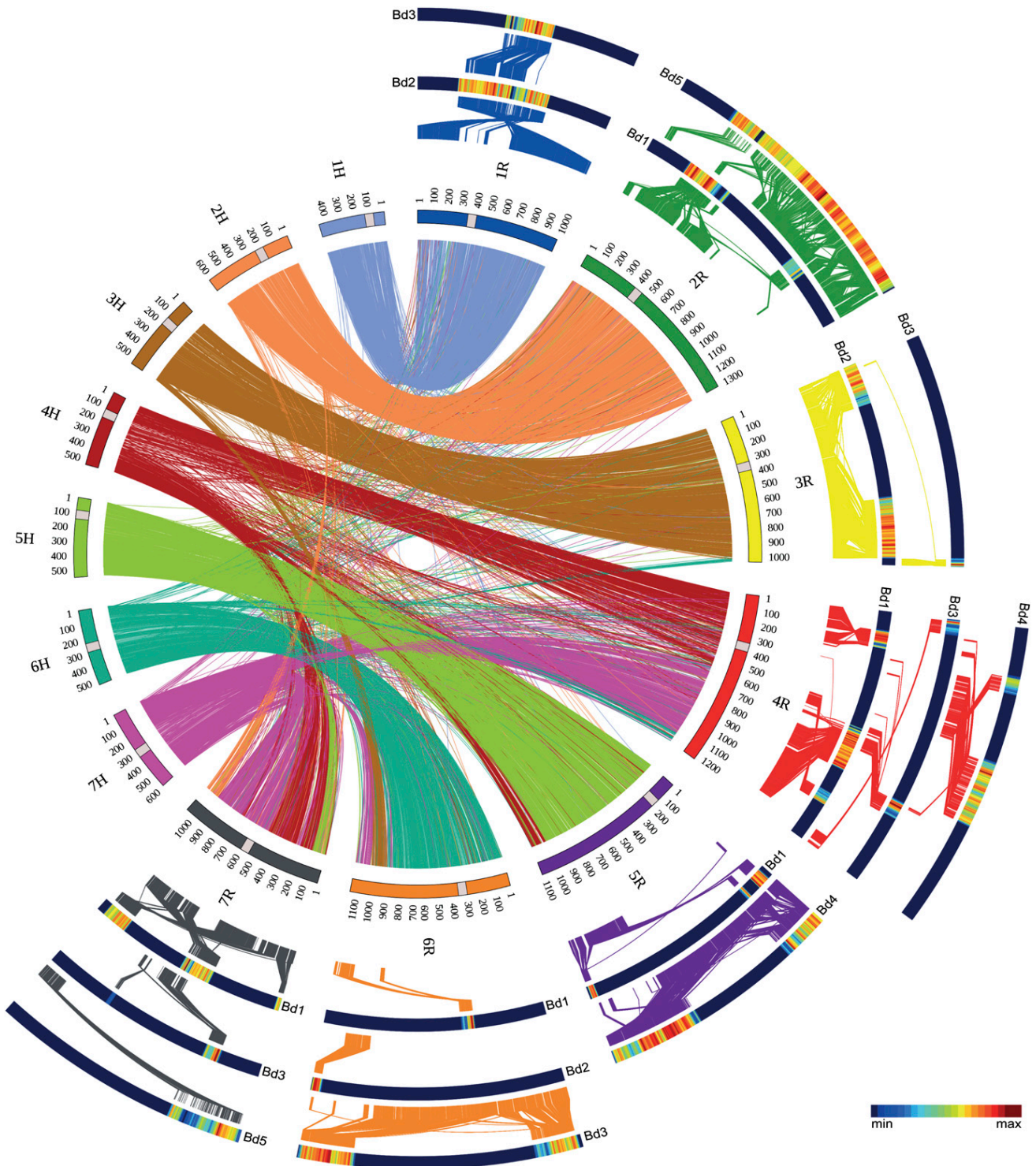


Figure 2. Conserved Synteny between Rye, Barley, and *B. distachyon*.

Collinearity of the rye and barley genomes is depicted by the inner circle of the diagram. Rye (1R to 7R) and barley (1H to 7H) chromosomes were scaled according to the rye genetic and barley physical map, respectively. Lines (colored according to barley chromosomes) within the inner circle connect putatively orthologous rye and barley genes. The outer partial circles of heatmap colored bars illustrate the density of *B. distachyon* genes hit by the

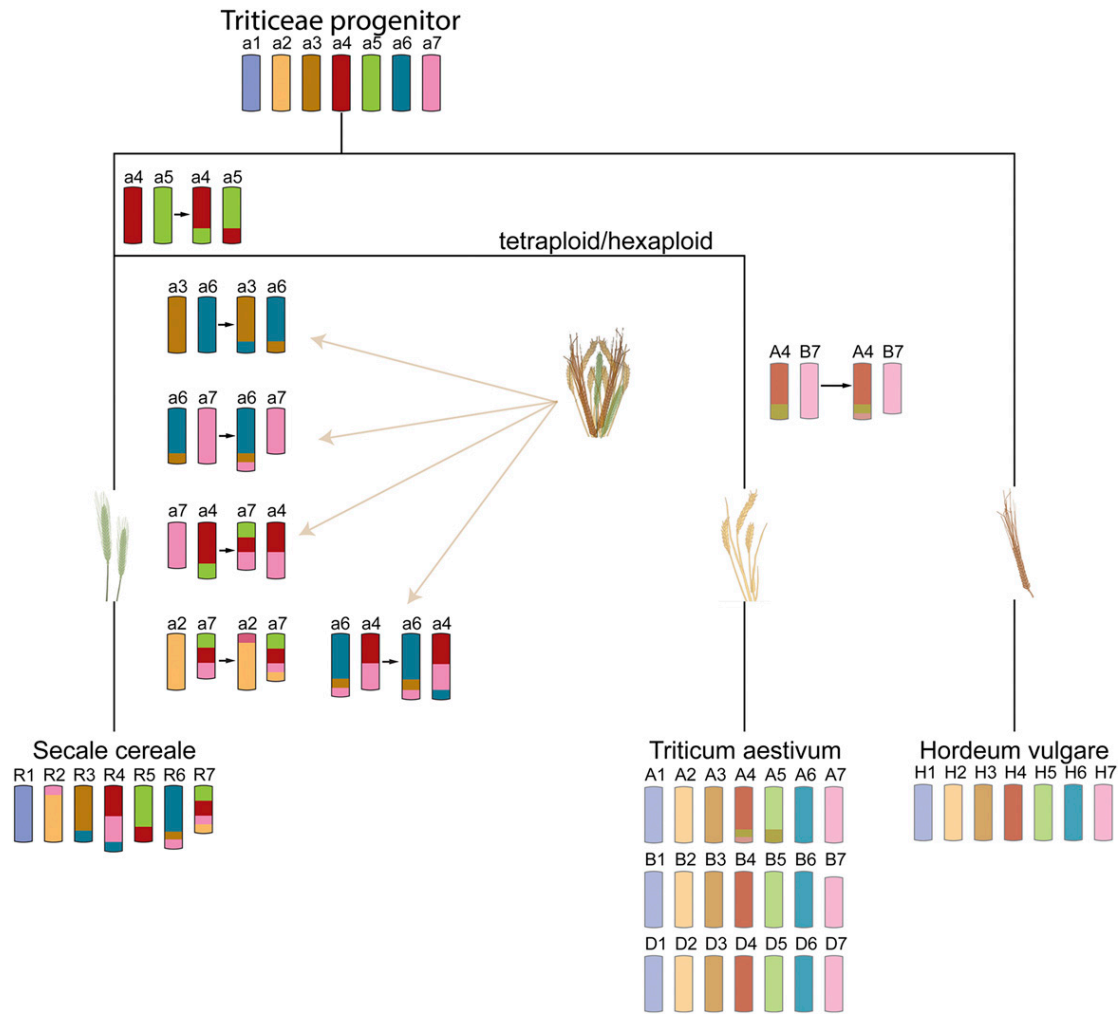


Figure 3. Rye Genome Reorganization and Translocation Events.

Rye genome reorganizations occurring in the common ancestor of rye and wheat (translocation between chromosomes 4 and 5) and divergence of the two lineages are postulated. Three of the five translocations that occurred after the split of wheat can be ordered, while for two the order cannot be deduced. They may have occurred in parallel or consecutively.

genes when compared with the three model genomes (Figure 4; see Supplemental Figure 5 online). Additionally, most segments contained also a similar fraction of conserved genes that were uniquely shared between either rye or barley and any of the three model genomes. However, four out of the 17 segments revealed pronounced deviations from this equilibrium. As an example, the distal conserved syntenic segment of chromosome 3R (denoted as 3R.2 in Figure 4) contained 10 to 16 times fewer conserved syntenic genes (30 to 48 genes) to *B. distachyon*, rice, and

sorghum than the putative orthologous segment of barley 6H (190 to 250 genes). Opposite examples were found for the most proximal segments of 7R (7R.4) or 4R (4R.1) (see Supplemental Figure 5 online) carrying up to 8 times more conserved syntenic genes to *B. distachyon*, rice, and sorghum than the respective segments of barley chromosomes 2H and 4H. The observed patterns could be due to differential retention of paralogs in rye and barley, differential evolutionary fate of conserved syntenic chromosome segments, or, in part, different evolutionary origins

Figure 2. (continued).

454 chromosome survey sequencing reads of the corresponding rye chromosomes. Conserved syntenic blocks are highlighted by yellow-red-colored regions of the heat maps. Putatively orthologous genes between rye and *B. distachyon* are connected with lines (colored according to rye chromosomes), and centromere positions are highlighted by gray rectangles.

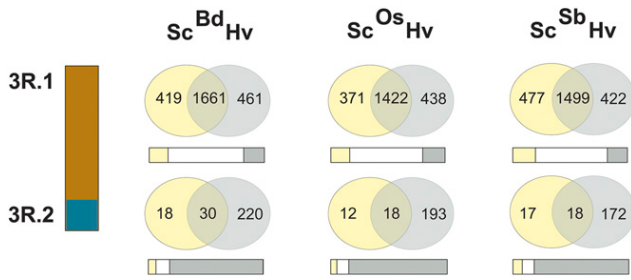


Figure 4. Conserved Synteny Statistics of Rye Chromosome 3R and the Corresponding Barley Regions to Reference Genomes.

Venn diagrams show the absolute number of conserved syntenic rye (yellow) and barley (gray) genes in comparison to the reference grass genomes of *B. distachyon*, rice, and sorghum. The bars below depict the percentage of distribution of reference genes shared by barley and rye (white), or rye (yellow) and barley alone (gray), respectively. While the 3R.1 fragment shows a balanced conserved syntenic pattern, the second fragment 3R.2 showed 10-fold less conserved syntenic genes in comparison to the corresponding barley segment.

of the corresponding segments and/or their parts. We found significant differences between the syntenic segments of rye and barley regarding the number of conserved syntenic genes for each of the three reference genomes (Pearson's χ^2 test; 32 *df*; $P < 1 \times 10^{-6}$).

Varying Sequence Identity Thresholds in Conserved Syntenic Segments Indicate Reticulate Evolution of the Rye Genome

The observation of unbalanced conserved syntenic gene content of orthologous genome segments of rye and barley in comparison to model grasses prompted us to expand our analysis toward testing for sequence conservation of the involved genes. We assessed sequence conservation of all anchored genomic sequence reads assigned to the 17 rye genome segments against a set of 28,622 full-length cDNAs (fl-cDNAs) of barley (Matsumoto et al., 2011). Corresponding orthologous genes and gene segments were selected using a first best hit criterion, and matching sequence regions had to exceed 100 nucleotides (≥ 30 amino acids). We plotted the sequence identity distribution for the 17 rye genomic fragments as heat map distributions (Figure 5A) and performed hierarchical clustering including 10,000-fold bootstrap resampling of sequence identity distributions for the respective segments. A broad distribution of sequence identity profiles was observed. Many segments (7R.3, 5R.1, 6R.1, 3R.1, 1R.1, 2R.2, and 4R.1) revealed overall sequence similarity in a relatively narrow range grouped around a maximum at 95% sequence identity. However, several individual segments (e.g., 2R.1, 3R.2, 6R.2, 6R.3, 4R.3, and 7R.4) exhibited a significant shift toward lower maximum sequence identity (Figure 5A). Statistical significance of sequence identity values was tested for segment-specific distributions also considering the amount of genes in the respective segment using a permutation test. For segment 2R.1, results were inconclusive, similar to previous results from the bootstrap clustering, most likely due to its small size. Strikingly, most segments

involved in rye lineage specific translocations (Figures 3 and 5) showed deviating identity profiles and grouped more distantly by hierarchical clustering (Figure 5B).

We expanded this analysis and measured synonymous (K_s) and nonsynonymous (K_a) substitution rates between rye/barley orthologs that were identified in the 17 conserved syntenic genome segments (see Supplemental Figure 6 online). Similar to the findings reported above, chromosomes 2R to 7R, all of which are composed of different syntenic segments with respect to barley, showed heterogeneous K_s mean and median values. The K_s distribution between the groups was significantly different (Kruskal-Wallis-test; $P < 0.004351$). However K_a/K_s values for the individual segments did not reveal pronounced differences; hence, no pattern of potential positive selection on individual genomic segments could be observed that might have caused the pronounced shifts in sequence similarities found for the individual rye segments.

Phylogenetic Analysis of Rye Chromosome Segments Indicates Variable Phylogenetic Networks

In a subsequent step, we analyzed the similarities and differences in phylogenetic networks for the 17 syntenic segments found in the rye genome. For each segment, we selected corresponding genes from five grass genomes for which either complete or draft genome sequences in different depth and resolution are available. Besides the rice genome that served as an outgroup, we also used

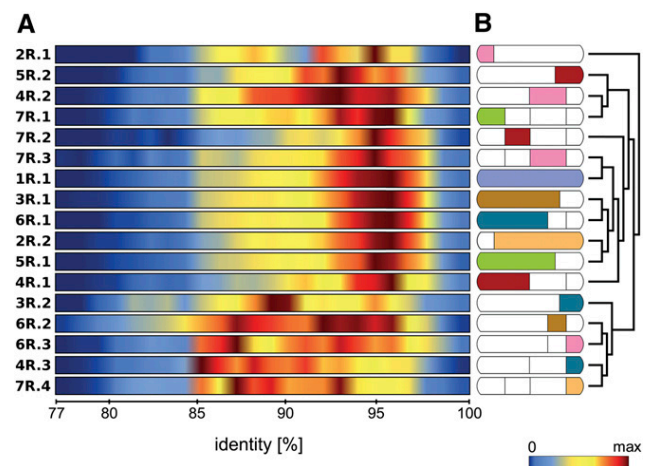


Figure 5. Sequence Conservation between Rye and Barley in 17 Conserved Syntenic Genome Segments.

(A) Rye gene-based chromosome survey sequences of the 17 conserved syntenic genome segments were compared with the putative barley orthologs (on the basis of fl-cDNAs) and the distribution of percentage of sequence identity is depicted by heat maps for each conserved block (max = highest no. of reads per segment with the given identity value; each block has its own maximum). The segments showed nonuniform sequence conservation patterns.

(B) The obtained sequence identity values were grouped by hierarchical clustering (average linkage, Euclidean distance) with the aim to find similarities between segments that could indicate their origin from the same progenitor genome and translocation or introgression event.

the genome of *B. distachyon*, the barley genome, and the recently published genome sequences of the two diploid wheat sub-genome progenitor species *Aegilops tauschii* and *Triticum urartu* (Jia et al., 2013; Ling et al., 2013). Corresponding genes were selected using a bidirectional best BLAST hit criterion, and a total of 705 gene clusters were generated and analyzed for phylogenetic networks (see Supplemental Figure 7 online). This analysis revealed that, consistent with the clustering results obtained using sequence conservation (Figure 5), rye genomic segments group differently in the phylogenetic networks. For eight rye segments (1R.1, 2R.2, 3R.1, 4R.2, 5R.1, 6R.1, 7R.2, and 7R.3), results indicate phylogenetic positioning of rye between barley and the wheat lineage (*Ae. tauschii* and *T. urartu*), but for other segments, the network structure was different, with varying relationship differences (e.g., 4R.1 found to group distant from the Triticeae). In addition, even within segments we found evidence for reticulate evolution for several segments (4R.3, 5R.2, 6R.2, and 7R.1). Thus, in summary, the phylogenetic networks for the 17 rye segments showed pronounced differences and even within some of the segments evidence for reticulate evolution was found.

DISCUSSION

Rye Genome Unlocked by Chromosomal Genomics

Wheat, barley, and rye are very closely related cereal crop species that were domesticated during a very narrow time span during the Neolithic Era. Their domestication was of critical importance for the establishment of early civilizations of the Fertile Crescent area in Near East and the spread of agriculture to Europe and Asia. For understanding evolution and domestication of the three species, as well as for any molecular genomic crop improvement strategy, it is a prerequisite to have access to (complete) genome sequence information. Significant progress has recently been reported for barley (Mayer et al., 2011; International Barley Genome Sequencing Consortium, 2012), wheat (Brenchley et al., 2012), and diploid wheat progenitor species (Jia et al., 2013; Ling et al., 2013). In this study, the rye genome could be unlocked by a combined approach of chromosomal genomics and conserved synteny analysis, providing comprehensive access to gene content as well as linear gene order information of about two thirds of the predicted rye genes.

We adopted an *in silico* method to establish so-called genome zippers to develop virtual linear gene order models that comprise considerable proportions of the genes of the ~8-Gb rye genome. This advance delivered an enabling platform for future genome-based rye research and improvement but also for high-resolution comparative analysis of related Triticeae species and grass genomes in general. The procedure integrated gene content information with a dense genetic map and conserved synteny information provided by reference sequences of related model grass genomes. The method has been proven successful and powerful for barley (Mayer et al., 2011), *Lolium* (Pfeifer et al., 2013), and wheat chromosome 4A (Hernandez et al., 2012). We used DNA amplified from flow-sorted rye chromosomes to generate CSS data, and ~31,000 genes were detected by sequence comparisons. Based on the measured sensitivity, ~40,000 genes can be postulated for the entire rye genome. However, this

number might be overestimated since gene fragments and pseudogenes are abundant in Triticeae genomes (Mayer et al., 2011; Wicker et al., 2011; International Barley Genome Sequencing Consortium, 2012), and due to the limited sequence coverage of the presented data sets, conclusions about the total gene set remain preliminary. Overall, this number is higher than, but comparable to, previous gene counts reported for other Triticeae genomes and rye chromosomes (Mayer et al., 2011; Martis et al., 2012; International Barley Genome Sequencing Consortium, 2012), suggesting that haploid gene content is similar in rye, barley, and wheat. A total of 22,426 genes (72% of the postulated genes) could be integrated into the rye genome zippers on the basis of the newly developed high-density gene-based genetic map of rye and conserved synteny information of the sequenced genomes of *B. distachyon*, rice, and sorghum. This number is similar to previous work, which identified 21,766 genes using the genome zipper approach for barley (Mayer et al., 2011).

Genome Collinearity between Rye and Barley

Syntenicity of grass genomes has been intensively studied, starting about two decades ago, on the basis of comparative RFLP mapping. Grass genomes share extensively conserved synteny and a circular model to visualize collinearity between smaller (i.e., rice) and larger grass genomes (i.e., Triticeae) was introduced (Moore et al., 1995). This model has been repeatedly revised as higher density maps became available for individual species (Devos, 2005) and recently has been enriched for information on ancient whole-genome duplication events leading to a refined model of grass karyotype evolution (Murat et al., 2010). We used the rye genome zippers developed in this work to reassess Triticeae genome collinearity and identified 17 segments representing the rye genome and exhibiting conserved synteny to the barley genome (International Barley Genome Sequencing Consortium, 2012). Rye chromosome 1R was the only linkage group that was collinear over its entire length to a single barley chromosome (1H). All other rye chromosomes were composed of between two and four segments corresponding to individual regions on the barley genome. However, our findings largely confirm earlier studies at unprecedented density and resolution since previous descriptions relied on mapping of 150 RFLP markers (Devos et al., 1993) in comparison to wheat. The major patterns of rearrangement between rye and barley can be described as a series of six subsequent translocation events, which we illustrate in a revised model of rye genome evolution. Starting from a set of seven ancestral Triticeae chromosomes that most closely resemble in organization the modern barley (HH) and *Ae. tauschii* (DD) genomes, four translocation events in rye can be sequentially ordered while the succession of two additional events remains uncertain. The initial translocation between ancestral chromosomes a4 and a5 is very similar and possibly homologous to a reciprocal translocation reported for the 4A and 5A chromosomes of wheat (Naranjo et al., 1987; Liu et al., 1992). In this scenario, three subsequent translocations between the ancestral chromosomes a3 and a6, a6 and a7, and a7 and a4 would have occurred. The two remaining translocations (a2/a7 and a6/a4) have likely taken place after the

three preceding translocations. However, their sequential order remains unclear and both events may have occurred at the same time.

What Mechanisms Have Shaped the Modern Rye Genome?

The unprecedented access to rye genomic sequence information provided with this study as well as the detailed genome sequence information recently published for barley (International Barley Genome Sequencing Consortium, 2012) allowed a detailed comparative analysis of conserved orthologous genomic segments between both genomes. This revealed that individual conserved syntenic genomic segments of rye and barley carried strikingly different numbers of putatively conserved orthologous genes in comparison to the model grass genomes of rice, *B. distachyon*, and sorghum. Furthermore, the genes of defined conserved syntenic rye genome segments exhibited significantly different signatures of sequence conservation if compared with their putatively orthologous barley gene sequences.

Analysis of synonymous and nonsynonymous substitutions did not provide any evidence of different selective pressure among the different genomic regions of rye, but phylogenetic analysis of individual rye genomic segments revealed pronounced differences in their relationships to the five compared grass species. The observed network structures are largely consistent with the results obtained by comparison of global sequence similarities of genes found in specific genomic segments. For eight of the segments, the consensus tree/network structure positions rye between barley and the wheat lineage, but for the other segments, differing phylogenetic networks were found. It is noteworthy that patterns of reticulate evolution were found in four of the segments. Thus, overall, we conclude that the rye genome represents a concatenation of genomic segments with, in part, differing evolutionary origins. Hence, the rye genome, to some extent, was likely shaped by introgressive hybridization or reticulate evolution.

It is important to note that reticulate genome evolution was postulated recently for rye by a multigenic phylogeny analysis (one chloroplast gene, 26 nuclear genes) of different Triticeae species (Escobar et al., 2011). Reticulate evolution or hybrid speciation was postulated to have occurred frequently during plant evolution (Kellogg and Bennetzen, 2004; Linder and Rieseberg, 2004; Mallet, 2005). In the Triticeae, it may have occurred in diploid species (Kellogg et al., 1996; Escobar et al., 2011), but it has been most frequently postulated for allopolyploid Triticeae genera (Kellogg et al., 1996; Mason-Gamer, 2004; Mason-Gamer et al., 2010; Mahelka et al., 2011). Reticulate or hybrid speciation can occur (reviewed in Linder and Rieseberg, 2004) as a consequence of allopolyploidization, which involves fusion of unreduced gametes, or instant genome duplication after fusion of haploid gametes, giving rise to a fertile hybrid species in which diploid parental genomes are maintained. This mechanism has been documented in a number of taxa, including *Brassica* and *Triticum* (Snowdon, 2007; Feldman and Levy, 2012). Reticulate speciation can also occur by diploid (homoploid) hybrid speciation, which involves fusion of reduced gametes of parental species (reviewed in Rieseberg, 1997; Linder and Rieseberg, 2004). Allopolyploid formation had a major impact on wheat evolution and provided advantages to new plant species to colonize new

niches (Levy and Feldman, 2002; Matsuoka, 2011). Diploid hybrid species of sunflower (*Helianthus annuus*) exhibited a selective advantage over their parental species in more extreme habitats, as demonstrated by resynthesized hybrid species (Rieseberg et al., 2003). In the sedge species *Carex curvula*, it has been postulated that interspecies hybrid formation could have provided an advantage under changing environmental conditions (Choler et al., 2004). Furthermore, chromosomal aberrations and spontaneous aneuploidy were observed to occur at higher frequency in *Aegilops speltoides* populations in marginal environments (Belyayev and Raskina, 2013).

Whether allopolyploid or diploid hybrid speciation provided more likely mechanisms shaping the modern rye genome remains speculative. Given the diploid nature of today's rye, it seems more intuitive to propose that rye underwent one or more diploid hybrid speciation events. The obligate outbreeding nature of rye may support that diploid hybrid speciation played a role in rye evolution since there is a strong correlation between outcrossing and diploid hybrid speciation in plant species with a confirmed reticulate evolutionary history (reviewed in Rieseberg, 1997). In this study, we found no obvious evidence of the allopolyploid nature of the rye genome. We identified no traces of additional whole-genome duplication (data not shown), besides the one shared by rice and other Triticeae species (Salse et al., 2008; Thiel et al., 2009). However, in comparison to the closely related barley and wheat genomes, rye has a 50% bigger monoploid genome, and it carries the highest number of translocations in comparison to a postulated ancestral Triticeae progenitor genome. It is tempting to speculate that rye genome evolution involved one (or more) episode(s) of polyploidization and/or interspecific hybridization between as yet unknown species leading to allopolyploidization. Thus, modern rye genome structure with seven chromosomes would be the outcome of extensive karyotype repatterning and diploidization. Cytological studies of interspecific hybrids in the genus *Secale* indicated that cultivated rye differs by three reciprocal translocations from its putative wild ancestors (Stutz, 1972; Singh and Röbbelen, 1977). It was hypothesized that cultivated rye *S. cereale* evolved from *Secale vavilovii* possibly after multiple introgressions from *Secale montanum*/*Secale strictum*. This is consistent with the idea of reticulate evolution of the genome of *S. cereale* with multiple introgression events and could also explain the different levels of sequence homology to barley for the individual corresponding genomic segments. Reciprocal translocations in combination with dysploid chromosome number reduction could explain how rye returned to a diploid status with extensive collinearity to the present day diploid Triticeae genomes (mechanism reviewed in Schubert and Lysak, 2011). In this scenario, the increased monoploid genome size of rye and the slightly increased gene content in comparison to diploid barley and wheat genomes may represent remnants of the allopolyploid origin of rye. The presence of B chromosomes in rye provides more support for the hypothesis that interspecies hybridization played a role in rye genome evolution (B chromosomes are absent in barley and wheat). B chromosomes are supernumerary chromosomes that do not follow Mendelian inheritance and may origin from standard A chromosomes after interspecific hybridization (reviewed in Camacho et al., 2000); however, they may also form without the need of hybridization. Survey sequenced flow-sorted

rye B chromosomes carried thousands of gene signatures with homology to rye chromosomes 3R and 7R (Martis et al., 2012). Thus, rye B chromosomes can also be interpreted as side products of reorganization of the genome after hybridization or whole-genome duplication and subsequent rediploidization. In this scenario, the B chromosomes and their apparent correspondence to regions of the A genome can be seen as indicative for genomic segments that got eliminated from the A genome during the reshaping/diploidization process.

Outlook

Next-generation sequencing and chromosome flow sorting allowed us to greatly improve the genomic resources for rye genome analysis. This will facilitate future work toward molecular crop improvement as well as the more targeted characterization and utilization of genetic resources and crop wild relatives in rye breeding. The global analysis of conserved synteny and sequence conservation to related grass species provided a comprehensive novel insight into current state rye genome organization and indicates a history of the rye genome possibly involving reticulate evolution. With the recent relatively easy access to genome-wide sequence information, even from large genomes like those of the Triticeae, a much more fine-grained picture of grass species evolution can be expected for the near future that will provide us with novel insights into the dynamics of grass genome evolution over time.

METHODS

Plant Material

Four mapping populations, Lo7xLo225, P87xP105, Lo90xLo115, and L2039-NxDH, were employed for high-throughput genotyping. Lo7xLo225 was derived from an interpool cross between two inbred lines Lo7 and Lo225 by KWS LOCHOW, and 131 RILs (F4) from this cross were developed at the Julius Kühn-Institut. For P87xP105, 69 RIL F6 lines were derived from a pair of reciprocal crosses of the two inbred parents P87 and P105. The population was developed at the Institute of Genetics and Cytology, Minsk, Belarus, by T.S. Schilko (Korzun et al., 1998). For Lo90xLo115, 220 RIL F4 lines were obtained from a cross between two inbred lines Lo90 and Lo115 by KWS LOCHOW. For L2039-NxDH, 100 RIL F9 lines that originate from an interpool cross between an elite inbred nonrestorer inbred line (L2039-N source: HYBRO) as female parent and a doubled haploid (DH) recombinant line (L285xL290, developed at the University of Hohenheim, Germany) were established at the Julius Kühn-Institut.

Molecular Marker Resources

A custom rye (*Secale cereale*) 5k Illumina iSelect array comprising 5234 EST-derived SNP markers (Haseneyer et al., 2011) was used for high-throughput genotyping. Furthermore, 1385 gene-based SSRs were data-mined and evaluated for their use as SSR markers from previously published rye EST resources (Haseneyer et al., 2011) by applying the software tool Misa (Thiel et al., 2003). In addition, 45 more markers (SSR and STS) previously mapped in different rye populations (Hackauf et al., 2009, 2012) provided anchoring information to other published genetic maps of rye and to assign the obtained L2039-NxDH-linkage groups to the seven rye chromosomes and for orienting chromosome maps. The marker TC427 (ALDH2b) was derived from a rye mitochondrial aldehyde dehydrogenase mRNA sequence (GenBank accession number AB084896.1) and assayed

using the primer pair 5'-TGTCCTGGTTGAAAAACAG-3' and 5'-TGATGTATGGCTGGAAAGTTG-3' as previously described (Hackauf and Wehling, 2005).

SNP Genotyping and Data Processing

A total of 300 ng of genomic DNA per plant was used for genotyping on the Illumina iScan platform with the Infinium HD assay following manufacturer's protocols. The fluorescence images of an array matrix carrying Cy3- and Cy5- labeled beads were generated with the two-channel scanner. Raw hybridization intensity data processing, clustering, and genotype calling (AA, AB, and BB) were performed using the genotyping module in the GenomeStudio software V2009.1 (Illumina). Genotyping data were cleaned by excluding SNP markers with (1) a GenTrain score <0.6, (2) >10% missing data, or (3) monomorphic pattern.

Genotyping EST-Derived SSR Markers

A total of 688 EST-derived rye SSR markers were screened for polymorphism in four parents (Lo7, Lo90, Lo115, and Lo225) of two mapping populations (Lo7xLo225 and Lo90xLo115). The respective progenies were genotyped with 271 polymorphic markers. PCR was conducted in a total volume of 20 μ L (20 ng of genomic DNA, 1 \times HotStar Taq PCR buffer, 250 nM each primer, 200 μ M deoxynucleotide triphosphates, and 0.5 units of HotStar Taq DNA polymerase [Qiagen]). A touch-down PCR profile was applied (initial denaturation: 15 min at 95°C, 45 cycles: denaturation at 94°C for 1 min, annealing for 1 min [1°C incremental reduction from 65 to 55°C in the first 10 cycles and then 55°C] and extension at 72°C for 1 min [10 min at final extension]). PCR products were resolved on 1.5% agarose gels. Only markers with <10% missing values were used for mapping. Primer sequences of 688 tested and 271 mapped EST-SSRs are given in Supplemental Data Set 8 online.

Construction of Individual and Consensus Linkage Maps

Map construction of populations Lo7xLo225, L2039-NxDH, and Lo90xLo115 was performed with JoinMap 4.0 (Kyazma). Grouping was performed at an independence logarithm (base 10) of odds score between 4.0 and 10.0. For locus ordering, the maximum likelihood algorithm was used. The genetic linkage map of the P87xP105 population was constructed using MSTMap (Wu et al., 2008) at the probability level $1E^{-7}$. The centimorgan distances were calculated by applying the Kosambi mapping function (Kosambi, 1944). In populations Lo7xLo225 and Lo90xLo115, SSR markers were distributed manually to the SNP-based linkage maps using the software MapManager QTX (Manly et al., 2001).

A draft consensus map based on the four individual linkage maps was constructed using MergeMap (Wu et al., 2008). The consensus linkage groups were then compared with the original four homologous linkage groups in order to identify conflicts in marker order. MapChart v2.2 (Voorrips, 2002) and Circos (Krzywinski et al., 2009) were used for graphical representation of the linkage maps. Genotyping and detailed map information of the individual and the consensus map are provided as Supplemental Data Sets 9 and 10 online.

Purification and Amplification of Chromosomal DNA for Sequencing

Aqueous suspensions of intact mitotic chromosomes were prepared from root tips of seedlings ('Imperial' rye for 1R and 'Chinese Spring'-'Imperial' wheat [*Triticum aestivum*]-rye disomic chromosome addition lines for 2R to 7R; Driscoll and Sears, 1971), and rye chromosomes 1R to 7R were purified using FACSAria SORP flow sorter (BD Biosciences) as described earlier (Kubaláková et al., 2003). Approximately 20,000 copies of each rye chromosome were flow-sorted, and their DNA was purified and multiple-

displacement amplified (MDA) by the Illustra GenomiPhi V2 DNA amplification kit (GE Healthcare) in three independent reactions as described before (Simková et al., 2008). MDA DNA samples from each chromosome were pooled prior to sequencing. The identity and purity of sorted chromosome fractions was determined using fluorescence in situ hybridization with pSc119.2 and 5S rDNA probes (Kubaláková et al., 2003) (see Supplemental Figures 8 and 9 online). The purity of flow-sorted chromosome fractions and resulting quantities of amplified chromosomal DNA are summarized in Supplemental Table 3 online.

Roche/454 Sequencing

DNA amplified from sorted chromosomes was used for Roche/454 shotgun sequencing. Five micrograms of individual chromosome MDA DNAs was used to prepare the 454 sequencing libraries with the GS Titanium General Library Preparation Kit following the manufacturer's instructions (Roche Diagnostics). The 454 sequencing libraries were processed utilizing the GS FLX Titanium LV emPCR (Lib-L) and GS FLX Titanium Sequencing (XLR70) kits (Roche Diagnostics) according to the manufacturer's instructions. Statistics and details about the CSS data are summarized in Table 2 and Supplemental Table 1 online. Base pair coverage per chromosome was calculated according to Lander and Waterman (1988). The estimated values were tested by comparing the CSS data sets against the available genetically anchored sequence markers. The specificity (Sp) of individual rye chromosome data sets was determined as the proportion of false positive (FP) and true negative (TN) sequence matches with genetically anchored markers providing the reference ($Sp = \frac{n_{TN}}{n_{TN} + n_{FP}}$).

Bioinformatic Analyses: Identification of Repetitive Regions

The repetitive DNA content of CSS data was detected using Vmatch (<http://www.vmatch.de>) against the Munich Information Center for Protein Sequences-REdat Poaceae 8.6.2 repeat library (Nussbaumer et al., 2013). The following parameters were applied: 70% identity cutoff, 100-bp minimal length, seed length 14, exdrop 5, and e-value 0.001.

Analysis of Conserved Synteny

To assess the number of genes present in rye and to determine conserved syntenic regions between rye, barley (*Hordeum vulgare*; International Barley Genome Sequencing Consortium, 2012), and the three model grass genomes rice (*Oryza sativa*; International Rice Genome Sequencing Project, 2005), sorghum (*Sorghum bicolor*; Paterson et al., 2009), and *Brachypodium distachyon* (International Brachypodium Initiative, 2010), the repeat-filtered 454 sequence reads (with stretches of at least 100-bp nonmasked nucleotides) were compared against the protein sequences of the other grass species using BLASTX. Only homologs with at least 85% (barley), 75% (*B. distachyon*), or 70% (rice and sorghum) similarity and a minimum length of 30 amino acids were considered. Genes with multiple evidence were counted only once. The number of conserved genes was calculated using a sliding window approach (window size of 0.5 Mb; window shift of 0.1 Mb) and visualized by Circos heat maps (Krzywinski et al., 2009).

Generation of Rye Genome Zippers

Genetic map data, chromosomal gene content of rye, and conserved synteny information to model grass genomes were used for developing virtual gene order maps (genome zippers) of all seven rye chromosomes according to the earlier described approach (Mayer et al., 2011). This framework was substantiated by information based on rye EST assemblies (Haseneyer et al., 2011) and barley full-length cDNAs (Matsumoto et al., 2011). The genome zipper integration data sets are available as Supplemental Data Sets 1 to 7 online.

Analysis of Rye/Barley Synteny

The 2940 genetic markers of rye were compared via bidirectional BLASTN against 2785 genetic markers of barley (Close et al., 2009), and the homologous pairs were displayed in a scatterplot using matplotlib (Hunter, 2007). This comparison revealed syntenic segments and various chromosomal rearrangements. The same overall but higher density picture was obtained comparing the nonmasked 454 reads of the rye genome zippers against the physical/genetic barley genome scaffold (International Barley Genome Sequencing Consortium, 2012). The comparison was achieved using BLASTN (Altschul et al., 1990) with (1) the best match with minimum 85% identity and (2) a minimal alignment length of 100 bp. Subsequently, the conserved syntenic regions were detected using a sliding window approach (window size of 5 Mb; window shift of 1 Mb) and visualized by heat maps for each rye chromosome separately. The rye/barley orthologous pairs were defined using bidirectional BLASTN hits with the cutoff values mentioned above and plotted with the help of Circos (Krzywinski et al., 2009).

Assessment of Sequence Diversity and Conservation in Rye/Barley Conserved Syntenic Regions of the Rye Genome in Comparison to Other Grass Species

After manual inspection of the syntenic patterns between rye and barley, several distinct syntenic regions with a variable amount of reads (326 to 21,175) and genes (55 to 2,140) were defined. In the next step, these individual fragments were assigned to the virtual gene maps of barley and rye by investigating the rye reads and corresponding barley genes and their position in the genome zipper. To calculate the synonymous (K_s) and nonsynonymous (K_a) substitution rates between barley and rye, the 454 reads of the individual syntenic blocks were compared against the derived protein sequence from barley fl-cDNAs. The protein sequences of the barley fl-cDNAs were predicted using OrfPredictor (Min et al., 2005). The comparison and identification of protein alignments were done using BLASTX. All first best hits with at least 85% identity and a minimum of 50 amino acids without internal stop codon were filtered for further analysis. The K_a/K_s substitution rate was calculated using the YN00 module of the PAML 4 suite (Yang, 2007). In a last step, the average K_a and K_s values were calculated for those proteins that were tagged by multiple 454 reads. All K_s values up to 10 were used for statistical analysis. The K_s and K_a values were visualized by boxplots using the matplotlib library (MATLAB; MathWorks).

To test the sequence diversity in the syntenic fragments, the 454 reads assigned to the corresponding regions were compared using BLASTN against barley fl-cDNAs (28,622 sequences) (Matsumoto et al., 2011). The obtained sequence identities of all matches with at least 100-bp alignment length were summarized in bins and plotted. The individual blocks on particular chromosomes showed nonuniform distribution patterns. To group fragments with similar distribution, a hierarchical clustering of the identity bins was performed. We applied a hierarchical clustering, employing the Euclidean distance and average linkage.

Statistical Analysis

The syntenic conservation of both rye and barley against the three reference genomes (*B. distachyon*, rice, and sorghum) was tested for homogeneity with respect to the degree of syntenic conservation for each segment. For each reference organism, Pearson's χ^2 test was applied separately by comparing the numbers of barley and rye genes mapped against the reference across all syntenic fragments.

The significance of the identity values clustering was assessed using bootstrap resampling (B = 10,000) as implemented in the pvclust package in R (Suzuki and Shimodaira, 2006). The reported approximately unbiased P values indicate the significance of the observed cluster, with values close to 100 showing clusters that have the strongest support. As the segment size varied strongly (326 to 21,175), we tested whether the observed

patterns were random by employing a permutation test. For each syntenic segment (sample size N), we randomly drew N identity values from the complete set of identity values and tested whether these were significantly different from the observed values using a Kolmogorov-Smirnov test (Massey, 1951). This was repeated 10,000 times. These analyses were performed using R (<http://www.R-project.org>).

Differences between rye and barley distributions of the synonymous substitution rate (K_s) were tested with the Kruskal-Wallis test using the R software package (<http://www.R-project.org>).

Phylogenetic Analysis

To test for reticulate evolution/introgressive hybridization, the protein sequences of six distinct species (rye, barley, *Aegilops tauschii*, *Triticum urartu*, *B. distachyon*, and rice) that map to the 17 syntenic conserved regions were analyzed. For each segment, corresponding orthologous genes from the respective species were extracted using a bidirectional best BLAST hit criteria against the respective rye genes. To generate sufficient data points for all segments, either clusters of six corresponding genes (from rye, barley, rice, *B. distachyon*, *Ae. tauschii*, and *T. urartu*) or clusters of five corresponding genes (as before but without a corresponding gene from *T. urartu*) were extracted. A total of 705 gene clusters were generated. For each segment, the amount of gene clusters used varied between 1 and 160. The sequences of each cluster were aligned using MUSCLE (Edgar, 2004). The maximum likelihood phylogeny inference was constructed using FastTree2 (Price et al., 2010) with the JTT+CAT substitution model and the Shimodaira-Hasegawa test to compute the confidence values of tree branches. The trees were rooted by defining rice as outgroup. The level-k network consensus algorithm implemented in Dendroscope3 (Huson and Scornavacca, 2012) was used to combine and visualize the phylogenetic trees for each individual fragment into a single phylogenetic consensus network. Each network represents all clusters from all input trees, if the clusters appear in more than 30%.

Accession Numbers

Sequence data from this article were submitted to the European Bioinformatics Institute sequence read archive under study accession ID ERP001745, sample IDs ERS167396 to ERS167402, experiment IDs ERX140512 to ERX140518, run IDs ERR164635 to ERR164641.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Rye Consensus Transcript Map.

Supplemental Figure 2. Conserved Homologous Regions between Rye and Barley.

Supplemental Figure 3. Conserved Synteny between Rye, Barley, and Rice.

Supplemental Figure 4. Conserved Synteny between Rye, Barley, and Sorghum.

Supplemental Figure 5. Conserved Synteny between Rye and Barley with *B. distachyon*, Rice, and Sorghum Genomes.

Supplemental Figure 6. Sequence Conservation of Rye and Barley Genes in Corresponding Genome Segments.

Supplemental Figure 7. Phylogenetic Networks for Individual Segments of the Rye Genome.

Supplemental Figure 8. Flow Cytometric Sorting of Rye Chromosome 1R from cv Imperial.

Supplemental Figure 9. Example of the Use of Wheat-Rye Chromosome Addition Lines to Purify Chromosomes 2R to 7R Using Flow Sorting.

Supplemental Table 1. Sequence and Repeat Analysis Statistics for Individual Rye Chromosomes.

Supplemental Table 2. Genome Zipper Statistics for Rye/Barley Orthologous Genome Segments.

Supplemental Table 3. Purity of Flow-Sorted Rye Chromosome Fractions and DNA Amounts Obtained after Amplification of Chromosomal DNA.

Supplemental Data Set 1. Genome Zipper of Rye Chromosome 1R.

Supplemental Data Set 2. Genome Zipper of Rye Chromosome 2R.

Supplemental Data Set 3. Genome Zipper of Rye Chromosome 3R.

Supplemental Data Set 4. Genome Zipper of Rye Chromosome 4R.

Supplemental Data Set 5. Genome Zipper of Rye Chromosome 5R.

Supplemental Data Set 6. Genome Zipper of Rye Chromosome 6R.

Supplemental Data Set 7. Genome Zipper of Rye Chromosome 7R.

Supplemental Data Set 8. EST-Derived Rye SSR Markers.

Supplemental Data Set 9. Mapping Data of Four Populations.

Supplemental Data Set 10. Rye Consensus Transcript Map.

ACKNOWLEDGMENTS

We thank Adam Lukaszewski for providing seeds of rye cv 'Imperial' and wheat-rye chromosome addition lines, Jarmila Čihalíková, Zdenka Dubská, and Romana Šperková for assistance with chromosome sorting and DNA amplification, and Heidrun Gundlach for help in repeat masking. We also thank Bjoern Usadel and Doreen Pahlke from Plant2030-PD for support with submission of sequence data sets to the European Bioinformatics Institute. This work was financially supported by the following grants: GABI Barlex 0314000 to N.S. and K.F.X.M.; GABI Rye-Express 0315063 from the German Ministry of Education and Research (BMBF) to N.S., K.F.X.M., and E.B.; FP7-212019 TriticeaeGenome from the European Union commission to N.S., K.F.X.M., and J.D.; SFB 924 grant of the Deutsche Forschungsgemeinschaft to K.F.X.M.; and Czech Science Foundation Award P501/12/G090 and the Ministry of Education, Youth, and Sports of the Czech Republic and the European Regional Development Fund (Operational Programme Research and Development for Innovations No. ED0007/01/01) to J.D., M.K., and J.V.

AUTHOR CONTRIBUTIONS

K.F.X.M., E.B., and N.S. designed the research. R.Z., G.H., S.K., B.H., V.K., M.K., and J.V. performed experiments. E.B., G.H., B.H., V.K., T.S., and U.S. contributed data sets and analytical/computational tools. M.M.M., G.H., R.Z., K.G.K., and T.S. performed data analysis. K.F.X.M., M.M.M., R.Z., G.H., C.-C.S., E.B., J.D., and N.S. wrote/edited the article. All authors read and approved the article.

Received June 5, 2013; revised August 23, 2013; accepted September 20, 2013; published October 8, 2013.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.

- Belyayev, A., and Raskina, O.** (2013). Chromosome evolution in marginal populations of *Aegilops speltoides*: Causes and consequences. *Ann. Bot. (Lond.)* **111**: 531–538.
- Brenchley, R., et al.** (2012). Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature* **491**: 705–710.
- Camacho, J.P.M., Sharbel, T.F., and Beukeboom, L.W.** (2000). B-chromosome evolution. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **355**: 163–178.
- Choler, P., Erschbamer, B., Tribsch, A., Gielly, L., and Taberlet, P.** (2004). Genetic introgression as a potential to widen a species' niche: Insights from alpine *Carex curvula*. *Proc. Natl. Acad. Sci. USA* **101**: 171–176.
- Close, T.J., et al.** (2009). Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics* **10**: 582.
- Devos, K.M.** (2005). Updating the 'crop circle'. *Curr. Opin. Plant Biol.* **8**: 155–162.
- Devos, K.M., Atkinson, M.D., Chinoy, C.N., Francis, H.A., Harcourt, R.L., Koebner, R.M.D., Liu, C.J., Masojc, P., Xie, D.X., and Gale, M.D.** (1993). Chromosomal rearrangements in the rye genome relative to that of wheat. *Theor. Appl. Genet.* **85**: 673–680.
- Doležel, J., Bartoš, J., Voglmayr, H., and Greilhuber, J.** (2003). Nuclear DNA content and genome size of trout and human. *Cytometry A* **51**: 127–128, author reply 129.
- Doležel, J., Greilhuber, J., Lucretti, S., Meister, A., Lysák, M.A., Nardi, L., and Obermayer, R.** (1998). Plant genome size estimation by flow cytometry: Inter-laboratory comparison. *Ann. Bot. (Lond.)* **82**: 17–26.
- Doležel, J., Vrána, J., Safár, J., Bartoš, J., Kubaláková, M., and Simková, H.** (2012). Chromosomes in the flow to simplify genome analysis. *Funct. Integr. Genomics* **12**: 397–416.
- Driscoll, C., and Sears, E.** (1971). Individual addition of the chromosomes of 'Imperial' rye to wheat. *Agronomy Abstracts* **6**.
- Edgar, R.C.** (2004). MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**: 1792–1797.
- Escobar, J.S., Scornavacca, C., Cenci, A., Guilhaumon, C., Santoni, S., Douzery, E.J., Ranwez, V., Glémin, S., and David, J.** (2011). Multigenic phylogeny and analysis of tree incongruences in Triticeae (*Poaceae*). *BMC Evol. Biol.* **11**: 181.
- Feldman, M., and Levy, A.A.** (2012). Genome evolution due to allopolyploidization in wheat. *Genetics* **192**: 763–774.
- Geiger, H., and Miedaner, T.** (2009). Rye breeding. In *Handbook of Plant Breeding: Cereals*, M. Carena, ed (New York: Springer Science + Business Media), pp. 157–181.
- Gustafson, J.P., Ma, X.-F., Korzun, V., and Snape, J.W.** (2009). A consensus map of rye integrating mapping data from five mapping populations. *Theor. Appl. Genet.* **118**: 793–800.
- Hackauf, B., Korzun, V., Wortmann, H., Wilde, P., and Wehling, P.** (2012). Development of conserved ortholog set markers linked to the restorer gene *Rfp1* in rye. *Mol. Breed.* **30**: 1507–1518.
- Hackauf, B., Rudd, S., van der Voort, J.R., Miedaner, T., and Wehling, P.** (2009). Comparative mapping of DNA sequences in rye (*Secale cereale* L.) in relation to the rice genome. *Theor. Appl. Genet.* **118**: 371–384.
- Hackauf, B., and Wehling, P.** (2005). Approaching the self-incompatibility locus Z in rye (*Secale cereale* L.) via comparative genetics. *Theor. Appl. Genet.* **110**: 832–845.
- Haseneyer, G., Schmutzer, T., Seidel, M., Zhou, R., Mascher, M., Schön, C.C., Taudien, S., Scholz, U., Stein, N., Mayer, K.F., and Bauer, E.** (2011). From RNA-seq to large-scale genotyping - Genomics resources for rye (*Secale cereale* L.). *BMC Plant Biol.* **11**: 131.
- Hernandez, P., Martis, M., Dorado, G., Pfeifer, M., Gálvez, S., Schaaf, S., Jouve, N., Šimková, H., Valárik, M., Doležel, J., and Mayer, K.F.X.** (2012). Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant J.* **69**: 377–386.
- Huang, S., Sirikhachornkit, A., Faris, J.D., Su, X., Gill, B.S., Haselkorn, R., and Gornicki, P.** (2002). Phylogenetic analysis of the acetyl-CoA carboxylase and 3-phosphoglycerate kinase loci in wheat and other grasses. *Plant Mol. Biol.* **48**: 805–820.
- Hunter, J.** (2007). Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**: 90–95.
- Huson, D.H., and Scornavacca, C.** (2012). Dendroscope 3: An interactive tool for rooted phylogenetic trees and networks. *Syst. Biol.* **61**: 1061–1067.
- International Barley Genome Sequencing Consortium; Mayer, K.F., Waugh, R., Brown, J.W., Schulman, A., Langridge, P., Platzer, M., Fincher, G.B., Muehlbauer, G.J., Sato, K., Close, T.J., Wise, R.P., and Stein, N.** (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* **491**: 711–716.
- International Brachypodium Initiative** (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature* **463**: 763–768.
- International Rice Genome Sequencing Project** (2005). The map-based sequence of the rice genome. *Nature* **436**: 793–800.
- Jia, J., et al; International Wheat Genome Sequencing Consortium** (2013). *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature* **496**: 91–95.
- Kellogg, E.A., Appels, R., and Mason-Gamer, R.J.** (1996). When gene trees tell different stories: The diploid genera of Triticeae. *Syst. Bot.* **21**: 312–347.
- Kellogg, E.A., and Bennetzen, J.L.** (2004). The evolution of nuclear genome structure in seed plants. *Am. J. Bot.* **91**: 1709–1725.
- Korzun, V., Malyshev, S., Kartel, N., Westermann, T., Weber, W.E., and Börner, A.** (1998). A genetic linkage map of rye (*Secale cereale* L.). *Theor. Appl. Genet.* **96**: 203–208.
- Kosambi, D.** (1944). The estimation of map distances from recombination values. *Ann. Eugen.* **12**: 172–175.
- Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A.** (2009). Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**: 1639–1645.
- Kubaláková, M., Valárik, M., Barto, J., Vrána, J., Cíhalíková, J., Molnár-Láng, M., and Doležel, J.** (2003). Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. *Genome* **46**: 893–905.
- Lander, E.S., and Waterman, M.S.** (1988). Genomic mapping by fingerprinting random clones: A mathematical analysis. *Genomics* **2**: 231–239.
- Levy, A.A., and Feldman, M.** (2002). The impact of polyploidy on grass genome evolution. *Plant Physiol.* **130**: 1587–1593.
- Linder, C.R., and Rieseberg, L.H.** (2004). Reconstructing patterns of reticulate evolution in plants. *Am. J. Bot.* **91**: 1700–1708.
- Ling, H.-Q., et al.** (2013). Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature* **496**: 87–90.
- Liu, C., Atkinson, M., Chinoy, C., Devos, K., and Gale, M.** (1992). Nonhomoeologous translocations between group 4, 5 and 7 chromosomes within wheat and rye. *Theor. Appl. Genet.* **83**: 305–312.
- Lundqvist, A.** (1956). Self-incompatibility in rye. I. Genetic control in the diploid. *Hereditas* **42**: 293–348.
- Mahelka, V., Kopecký, D., and Paštová, L.** (2011). On the genome constitution and evolution of intermediate wheatgrass (*Thinopyrum intermedium*: *Poaceae*, Triticeae). *BMC Evol. Biol.* **11**: 127.
- Mallet, J.** (2005). Hybridization as an invasion of the genome. *Trends Ecol. Evol. (Amst.)* **20**: 229–237.
- Manly, K.F., Cudmore, R.H., Jr., and Meer, J.M.** (2001). Map Manager QTX, cross-platform software for genetic mapping. *Mamm. Genome* **12**: 930–932.

- Martis, M.M., et al.** (2012). Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc. Natl. Acad. Sci. USA* **109**: 13343–13346.
- Mason-Gamer, R.J.** (2004). Reticulate evolution, introgression, and intertribal gene capture in an allohexaploid grass. *Syst. Biol.* **53**: 25–37.
- Mason-Gamer, R.J., Burns, M.M., and Naum, M.** (2010). Reticulate evolutionary history of a complex group of grasses: Phylogeny of *Elymus* StStHH allotetraploids based on three nuclear genes. *PLoS ONE* **5**: e10989.
- Massey, F.J.** (1951). The Kolmogorov-Smirnov test for goodness of fit. *J. Am. Stat. Assoc.* **46**: 68–78.
- Matsumoto, T., et al.** (2011). Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from 12 clone libraries. *Plant Physiol.* **156**: 20–28.
- Matsuoka, Y.** (2011). Evolution of polyploid *Triticum* wheats under cultivation: The role of domestication, natural hybridization and allopolyploid speciation in their diversification. *Plant Cell Physiol.* **52**: 750–764.
- Mayer, K.F.X., et al.** (2011). Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell* **23**: 1249–1263.
- Mayer, K.F.X., et al.** (2009). Gene content and virtual gene order of barley chromosome 1H. *Plant Physiol.* **151**: 496–505.
- Milczarski, P., Bolibok-Bragoszewska, H., Myśków, B., Stojalowski, S., Heller-Uszyńska, K., Góralaska, M., Brągoszewski, P., Uszyński, G., Kilian, A., and Rakoczy-Trojanowska, M.** (2011). A high density consensus map of rye (*Secale cereale* L.) based on DArT markers. *PLoS ONE* **6**: e28495.
- Min, X.J., Butler, G., Storms, R., and Tsang, A.** (2005). OrfPredictor: Predicting protein-coding regions in EST-derived sequences. *Nucleic Acids Res.* **33** (Web Server issue): W677–W680.
- Moore, G., Devos, K.M., Wang, Z., and Gale, M.D.** (1995). Cereal genome evolution. Grasses, line up and form a circle. *Curr. Biol.* **5**: 737–739.
- Murat, F., Xu, J.-H., Tannier, E., Abrouk, M., Guilhot, N., Pont, C., Messing, J., and Salse, J.** (2010). Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a source of plant evolution. *Genome Res.* **20**: 1545–1557.
- Naranjo, T., Roca, A., Gooicoechea, P., and Giraldez, R.** (1987). Am homoeology of wheat and rye chromosomes. *Genome* **29**: 873–882.
- Nussbaumer, T., Martis, M.M., Roessner, S.K., Pfeifer, M., Bader, K.C., Sharma, S., Gundlach, H., and Spannagl, M.** (2013). MIPS PlantsDB: A database framework for comparative plant genome research. *Nucleic Acids Res.* **41** (Database issue): D1144–D1151.
- Paterson, A.H., et al.** (2009). The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **457**: 551–556.
- Pfeifer, M., Martis, M., Asp, T., Mayer, K.F.X., Lübberstedt, T., Byrne, S., Frei, U., and Studer, B.** (2013). The perennial ryegrass GenomeZipper: Targeted use of genome resources for comparative grass genomics. *Plant Physiol.* **161**: 571–582.
- Price, M.N., Dehal, P.S., and Arkin, A.P.** (2010). FastTree 2—Approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**: e9490.
- Qi, L.L., et al.** (2004). A chromosome bin map of 16,000 expressed sequence tag loci and distribution of genes among the three genomes of polyploid wheat. *Genetics* **168**: 701–712.
- Rieseberg, L.H.** (1997). Hybrid origins of plant species. *Annu. Rev. Ecol. Syst.* **28**: 359–389.
- Rieseberg, L.H., Raymond, O., Rosenthal, D.M., Lai, Z., Livingstone, K., Nakazato, T., Durphy, J.L., Schwarzbach, A.E., Donovan, L.A., and Lexer, C.** (2003). Major ecological transitions in wild sunflowers facilitated by hybridization. *Science* **301**: 1211–1216.
- Salse, J., Bolot, S., Throude, M., Jouffe, V., Piegu, B., Quraishi, U.M., Calcagno, T., Cooke, R., Delseny, M., and Feuillet, C.** (2008). Identification and characterization of shared duplications between rice and wheat provide new insight into grass genome evolution. *Plant Cell* **20**: 11–24.
- Sato, K., Nankaku, N., and Takeda, K.** (2009). A high-density transcript linkage map of barley derived from a single population. *Heredity (Edinb)* **103**: 110–117.
- Schlegel, R., Melz, G., and Nestrowicz, R.** (1987). A universal reference karyotype in rye, *Secale cereale* L. *Theor. Appl. Genet.* **74**: 820–826.
- Schubert, I., and Lysak, M.A.** (2011). Interpretation of karyotype evolution should consider chromosome structural constraints. *Trends Genet.* **27**: 207–216.
- Sencer, H., and Hawkes, J.** (1980). On the origin of cultivated rye. *Biol. J. Linn. Soc. Lond.* **13**: 299–313.
- Simková, H., Svensson, J.T., Condamine, P., Hříbová, E., Suchánková, P., Bhat, P.R., Bartoš, J., Safár, J., Close, T.J., and Doležel, J.** (2008). Coupling amplified DNA from flow-sorted chromosomes to high-density SNP mapping in barley. *BMC Genomics* **9**: 294.
- Singh, R.J., and Röbbelen, G.** (1977). Identification by Giemsa technique of the translocations separating cultivated rye from three wild species of *Secale*. *Chromosoma* **59**: 217–225.
- Snowdon, R.J.** (2007). Cytogenetics and genome analysis in *Brassica* crops. *Chromosome Res.* **15**: 85–95.
- Stein, N., Prasad, M., Scholz, U., Thiel, T., Zhang, H., Wolf, M., Kota, R., Varshney, R.K., Perovic, D., Grosse, I., and Graner, A.** (2007). A 1,000-loci transcript map of the barley genome: New anchoring points for integrative grass genomics. *Theor. Appl. Genet.* **114**: 823–839.
- Stutz, H.** (1972). On the origin of cultivated rye. *Am. J. Bot.* **59**: 59–70.
- Suzuki, R., and Shimodaira, H.** (2006). Pvcust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**: 1540–1542.
- Thiel, T., Graner, A., Waugh, R., Grosse, I., Close, T.J., and Stein, N.** (2009). Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. *BMC Evol. Biol.* **9**: 209.
- Thiel, T., Michalek, W., Varshney, R.K., and Graner, A.** (2003). Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). *Theor. Appl. Genet.* **106**: 411–422.
- Voorrips, R.E.** (2002). MapChart: Software for the graphical presentation of linkage maps and QTLs. *J. Hered.* **93**: 77–78.
- Wicker, T., et al.** (2011). Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives. *Plant Cell* **23**: 1706–1718.
- Willcox, G.** (2005). The distribution, natural habitats and availability of wild cereals in relation to their domestication in the Near East: Multiple events, multiple centres. *Veget. Hist. Archaeobot.* **14**: 534–541.
- Wu, Y., Bhat, P.R., Close, T.J., and Lonardi, S.** (2008). Efficient and accurate construction of genetic linkage maps from the minimum spanning tree of a graph. *PLoS Genet.* **4**: e1000212.
- Yang, Z.** (2007). PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**: 1586–1591.

Reticulate Evolution of the Rye Genome

Mihaela M. Martis, Ruonan Zhou, Grit Haseneyer, Thomas Schmutzer, Jan Vrána, Marie Kubaláková, Susanne König, Karl G. Kugler, Uwe Scholz, Bernd Hackauf, Viktor Korzun, Chris-Carolin Schön, Jaroslav Doležel, Eva Bauer, Klaus F.X. Mayer and Nils Stein
Plant Cell; originally published online October 8, 2013;
DOI 10.1105/tpc.113.114553

This information is current as of June 10, 2015

Supplemental Data	http://www.plantcell.org/content/suppl/2013/10/08/tpc.113.114553.DC1.html
Permissions	https://www.copyright.com/ccc/openurl.do?sid=pd_hw1532298X&issn=1532298X&WT.mc_id=pd_hw1532298X
eTOCs	Sign up for eTOCs at: http://www.plantcell.org/cgi/alerts/ctmain
CiteTrack Alerts	Sign up for CiteTrack Alerts at: http://www.plantcell.org/cgi/alerts/ctmain
Subscription Information	Subscription Information for <i>The Plant Cell</i> and <i>Plant Physiology</i> is available at: http://www.aspb.org/publications/subscriptions.cfm

Bibliography

- [1] Paterson, A., Freeling, M., Tang, H. et al. Insights from the comparison of plant genome sequences. *Annual Review of Plant Biology*, 61:349–372, 2010.
- [2] Campbell, N.A. *Biologie*. Spektrum Akademischer Verlag, 2. korrigierter nachdruck edition, 2000.
- [3] Drinnan, A., Crane, P. and Hoot, S. Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). *Plant Systematics and Evolution Supplement* 8, 8:93–122, 1994.
- [4] Soltis, P. and Soltis, D. The origin and diversification of angiosperms. *American Journal of Botany*, 91(10):1614–1626, 2004.
- [5] Secretariat of the Convention on Biological Diversity. The convention on biological diversity plant conservation report: a review of progress in implementing the global strategy of plant conservation (GSPC), 2009. URL <http://www.cbd.int/gspc/pcr-report/default.shtml>.
- [6] Raven, P., Evert, R. and Eichorn, S. *Biology of plants*. New York: Worth, 1992.
- [7] Lynch, A., Barnes, R., Cambecèdes, J. et al. Genetic evidence that *Lomatia tasmanica* (*Proteaceae*) is an ancient clone. *Aust. J. Bot.*, 46:25–33, 1998.
- [8] Claros, M., Bautista, R., Guerrero-Fernández, D. et al. Why assembling plant genome sequences is so challenging. *Biology*, 1:439–459, 2012.
- [9] Bennett, M. and Leitch, I. Nuclear DNA amounts in angiosperms: targets, trends and tomorrow. *Ann. Bot.*, 107:467–590, 2011.
- [10] Pellicer, J., Fay, M. and Leitch, I. The large eukaryotic genome of them all? *Botanical Journal of the Linnean Society*, 164(1):10–15, 2010.
- [11] Heslop-Harrison, J. and Schwarzacher, T. Organisation of the plant genome in chromosomes. *The Plant Journal*, 66:18–33, 2011.
- [12] Bennett, M. and Leitch, I. Plant DNA C-values database (release 6.0, dec. 2012), 2012. URL <http://www.kew.org/cvalues/>.
- [13] Wang, W., Haberer, G., Gundlach, H. et al. The *Spirodela polyrhiza* genome reveals insights into its neotenus reduction fast growth and aquatic lifestyle. *Nat. Commun.*, 5, 2014.

BIBLIOGRAPHY

- [14] Murray, B. Nuclear DNA amounts in gymnosperms. *Ann. Bot.*, 82:13, 1998.
- [15] Johnson, M., Kenton, A., Bennett, M. et al. *Voanioala gerardii* has the highest known chromosome number in the monocotyledons. *Genome*, 32(2):328–333, 1989.
- [16] Khandelwal, S. Chromosome evolution in the genus *Ophioglossum* L. *Botanical Journal of the Linnean Society*, 102(3):205–217, 2008.
- [17] Murray, B., Friesen, N. and Heslop-Harrison, J. Molecular cytogenetic analysis of *Podocarpus* and comparison with other gymnosperm species. *Annals of Botany*, 89:483–489, 2002.
- [18] Bowers, J., Chapman, B., Rong, J. et al. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature*, 422:433–438, 2003.
- [19] Blanc, G. and Wolfe, K. Widespread paleopolyploidy in model plant species inferred from age distribution of duplicated genes. *Plant Cell*, 16:1667–1678, 2004.
- [20] Cannon, S., Sterck, L., Rombauts, S. et al. Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. *PNAS*, 103(40):14959–14964, 2006. doi: 10.1073/pnas.0603228103.
- [21] Soltis, D., Albert, V., Leebens-Mack, J. et al. Polyploidy and angiosperm diversification. *American Journal of Botany*, 96(1):336–348, 2009.
- [22] Adams, K. and Wendel, J. Polyploidy and genome evolution in plants. *Current Opinion in Plant Biology*, 8:135–141, 2005.
- [23] Leitch, I. and Bennett, M. Polyploidy in angiosperms. *Trends in Plant Science*, 2(12): 470–476, 1997.
- [24] Masterson, J. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. *Science*, 264(5157):421–424, 1994.
- [25] Wendel, J. Genome evolution in polyploids. *Plant Molecular Biology*, 42:225–249, 2000.
- [26] Vision, T., Brown, D. and Tanksley, S. The origins of genomic duplications in *Arabidopsis*. *Science*, 290(5499):2114–2117, 2000. doi: 10.1126/science.290.5499.2114.
- [27] Messing, J., Bharti, A., Karlowski, W. et al. Sequence composition and genome organization of maize. *PNAS*, 101:14349–14354, 2004.
- [28] Schlueter, J., Dixon, P., Granger, C. et al. Mining EST database to resolve evolution events in major crop species. *Genome*, 47:868–876, 2004.
- [29] Paterson, A., Bowers, J. and Chapman, B. Ancient polyploidy predating divergence of the cereals, and its consequences for comparative genomics. *PNAS*, 101(26):9903–9908, 2004.

- [30] Han, F., Fedak, G., Ouellet, T. et al. Rapid genomic changes in interspecific and intergeneric hybrids and allopolyploids of *Triticeae*. *Genome*, 46:716–723, 2003.
- [31] Bento, M., Pereira, H., Rocheta, M. et al. Polyploidization as a retraction force in plant genome evolution: sequence rearrangement in *Triticale*. *PLoS One*, 3(1):e1402, 2008.
- [32] Madlung, A., Masuelli, R., Watson, B. et al. Remodeling of DNA methylation and phenotypic and transcriptional changes in synthetic *Arabidopsis allotetraploids*. *Plant Physiology*, 129:733–746, 2002.
- [33] Shaked, H., Kashkush, K., Ozkan, H. et al. Sequence elimination and cytosine methylation are rapid and reproducible responses of the genome to wide hybridisation and allopolyploidy in wheat. *The Plant Cell*, 13:1749–1759, 2001.
- [34] Liu, B. and Wendel, J. Epigenetic phenomena and the evolution of plant allopolyploids. *Mol Phylogenet Evol*, 29:365–379, 2003.
- [35] Feldman, M., Liu, B., Segal, G. et al. Rapid elimination of low-copy DNA sequences in polyploid wheat: a possible mechanism for differentiation of homoeologous chromosomes. *Genetics*, 147:1381–1387, 1997.
- [36] Han, F., Fedak, G., Guo, W. et al. Rapid and repeatable elimination of a parental genome-specific repeat (*pGc1R-1a*) in newly synthesized wheat allopolyploids. *Genetics*, 170(3):1239–1245, 2005.
- [37] Pestsova, E., Goncharov, N. and Salina, E. Elimination of tandem repeat of telomeric heterochromatin during the evolution of wheat. *Theor Appl Genet*, 97(8):1380–1386, 1998.
- [38] Kashkush, K., Feldman, M. and Levy, A. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nature Genetics*, 33(1):102–106, 2003.
- [39] Hanada, K., Vallejo, V., Nobuta, K. et al. The functional role of Pack-MULEs in rice inferred from purifying selection and expression profile. *The Plant Cell*, 21:25–38, 2009.
- [40] Morgante, M., Brunner, S., Pea, G. et al. Gene duplication and exon shuffling by helitron-like transposons genegene intraspecies diversity in maize. *Nature Genetics*, 37:997–1002, 2005.
- [41] Jiang, N., Bao, Z., Zhang, X. et al. Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, 431:569–573, 2004.
- [42] Jiang, N., Ferguson, A., Slotkin, R. et al. Pack-mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *PNAS*, 108(4):1537–1542, 2011.
- [43] Bennetzen, J. Transposable elements, gene creation and genome rearrangement in flowering plants. *Current Opinion in Genetics & Development*, 15:621–627, 2005.

BIBLIOGRAPHY

- [44] Madlung, A., Tyagi, A., Watson, B. et al. Genomic changes in synthetic *Arabidopsis* polyploids. *Plant Journal*, 41(2):221–230, 2005.
- [45] Kashkush, K., Feldman, M. and Levy, A. Gene loss, silencing and activation in a newly synthesized wheat allotetraploid. *Genetics*, 160(4):1651–1659, 2002.
- [46] Roth, C., Rastogi, S., Arvestad, L. et al. Evolution after gene duplication: model, mechanism, sequences, systems, and organisms. *J. Exp. Zool. (Mol. Dev. Evol.)*, 306B: 58–73, 2007.
- [47] Byrne, K. and Wolfe, K. Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. *Genetics*, 175:1341–1350, 2007.
- [48] Hughes, A. The evolution of nonfunctional novel proteins after gene duplication. *Proc. R. Soc. Lond. B*, 256, 1994. doi: 10.1098/rspb.1994.0058.
- [49] Lynch, M. and Force, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, 154:459–473, 2000.
- [50] Rastogi, S. and Liberles, D. Subfunctionalization of duplicated genes as a transition state to neofunctionalization. *BMC Evolutionary Biology*, 5(28), 2005. doi: 10.1186/1471-2148-5-28.
- [51] Comai, L., Tyagi, A., Winter, K. et al. Phenotypic instability and rapid gene silencing in newly formed *Arabidopsis* allotetraploids. *Plant Cell*, 12:1551–67, 2000.
- [52] Kidwell, M. and Lisch, D. Transposable elements as sources of variation in animal and plants. *Proc Natl Acad Sci U S A.*, 94:7704–7711, 1997.
- [53] Marillonnet, S. and Wessler, S. Retrotransposon insertion into the maize *waxy* gene results in tissue-specific RNA processing. *Plant Cell*, 9:967–978, 1997.
- [54] Blanc, G. and Wolfe, K. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell*, 16:1679–1691, 2004.
- [55] Bennetzen, J., Coleman, C., Liu, R. et al. Consistent over-estimation of gene number in complex plant genomes. *Current Opinion in Plant Biology*, 7:732–736, 2004.
- [56] Lee, T., Tang, H., Wang, X. et al. PGDD: a database of gene and genome duplications in plants. *Nucleic Acids Research*, pages 1–7, 2012. doi: 10.1093/nar/gks1104. <http://chibba.agtec.uga.edu/duplication/>.
- [57] Feuillet, C. and Keller, B. High gene density is conserved at syntenic loci of small and large grass genomes. *Proc Natl Acad Sci U S A.*, 96:8265–8270, 1999.
- [58] Madlung, A. Polyploidy and its effect on evolutionary success: old questions revisited with new tools. *Heredity*, 110:99–104, 2013.

- [59] Comai, L. The advantageous and disadvantages of being polyploid. *Nature Genetics*, 6, 2005.
- [60] Gu, Z., Steinmetz, L., Gu, X. et al. Role of duplicated genes in genetic robustness against null mutations. *Nature*, 421:63–66, 2003.
- [61] Birchler, J., Yao, H., Chudalayandi, S. et al. Heterosis. *The Plant Cell*, 22:2105–2112, 2010.
- [62] Bennetzen, J., Ma, J. and Devos, K. Mechanisms of recent genome size variation in flowering plants. *Ann. Bot. (Lond)*, 95:127–132, 2005.
- [63] Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–815, 2000.
- [64] Wicker, T., Mayer, K., Gundlach, H. 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):1706–18, 2011.
- [65] Jurka, J., Kapitonov, V., Kohany, O. et al. Repetitive sequences in complex genomes: structure and evolution. *Annu. Rev. Genomics Hum. Genet.*, 8:241–259, 2007.
- [66] McClintock, B. The origin and behaviour of mutable loci in maize. *PNAS*, 36(6):344–355, 1950.
- [67] Bennetzen, J. and Wang, H. The contribution of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.*, 65(505-30), 2014.
- [68] Wicker, T., Sabot, F., Hua-Van, A. et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.*, 8:973–982, 2007.
- [69] Baucom, R., Estill, J., Chaparro, C. et al. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genetics*, 5:e1000732, 2009.
- [70] SanMiguel, P. and Bennetzen, J. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann Bot*, 82:37–44, 1998.
- [71] Piegu, B., Guyot, R., Picault, N. et al. Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research*, 16:1262–1269, 2006.
- [72] Estep, M., DeBarry, J. and Bennetzen, J. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity*, 110:194–204, 2012. doi: 10.1038/hdy.2012.99.
- [73] Devos, K., Brown, J. and Bennetzen, J. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. *Genome Research*, 12: 1075–1079, 2002.

BIBLIOGRAPHY

- [74] Hawkins, J., Proulx, S., Rapp, R. et al. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *PNAS*, 106(42):17811–17816, 2009.
- [75] International Rice Genome Sequencing Project, I. The map-based sequence of the rice genome. *Nature*, 436:793–800, 2005.
- [76] Bolger, M., Weisshaar, B., Scholz, U. et al. Plant genome sequencing - applications for crop improvement. *Current Opinion in Biotechnology*, 26:31–37, 2014.
- [77] Schatz, M., Witkowski, J. and McCombie, W. Current challenges in *de novo* plant genome sequencing and assembly. *Genome Biol.*, 13:243, 2012.
- [78] Hamilton, J. and Buell, C. Advances in plant genome sequencing. *The Plant Journal*, 70: 177–190, 2012.
- [79] International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, 463:763–768, 2010.
- [80] Paterson, A., Bowers, J., Bruggmann, R. et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*, 457:551–556, 2009.
- [81] The French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaplohexaploid in major angiosperm phyla. *Nature*, 449:463–468, 2007.
- [82] Huang, S., Li, R., Zhang, Z. et al. The genome of the cucumber, *Cucumis sativus* L. *Nature Genetics*, 41(12):1275–1284, 2009.
- [83] Potato Genome Consortium. Genome sequence and analysis of the tuber crop potato. *Nature*, 475:189–195, 2011.
- [84] Shulaev, V., Sargent, D., Crowhurst, R. et al. The genome of woodland strawberry (*Fragaria vesca*). *Nature Genetics*, 43:109–116, 2011.
- [85] The Tomato Genome Consortium. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, 485:635–641, 2012.
- [86] D’Hont, A., Denoeud, F., Aury, J. et al. The banana (*Musa acuminata*) genome and the evolution of monocotyledonous plants. *Nature*, 488:213–217, 2012.
- [87] Nystedt, B., Street, N.R., Wetterbom, A. et al. The Norway spruce genome sequence and conifer genome evolution. *Nature*, 497:579–584, 2013.
- [88] Guo, S., Zhang, J., Sun, H. et al. The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions. *Nature Genetics*, 45:51–58, 2013.
- [89] Varshney, R., Song, C., Saxena, R. et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. *Nature Biotechnology*, 31:240–246, 2013.

- [90] Meyers, L. and Levin, D. On the abundance of polyploids in flowering plants. *Evolution*, 60(6):1198–1206, 2006.
- [91] Feuillet, C., Leach, J., Rogers, J. et al. Crop genome sequencing: lessons and rationales. *Trends in Plant Science*, 16(2):77–88, 2011.
- [92] Gnerre, S., MacCallum, I., Przybylski, D. et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *PNAS*, 108(4):1513–1518, 2011.
- [93] Luo, R., Liu, B., Xie, Y. et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1:18, 2012.
- [94] Simpson, J., Wong, K., Jackman, S. et al. ABySS: a parallel assembler for short read sequence data. *Genome Research*, 19:1117–1123, 2009.
- [95] Earl, D., Bradnam, K., St. John, J. et al. Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research*, 21:2224–2241, 2011.
- [96] Bradnam, K., Fass, J., Alexandrov, A. et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2:10, 2013. doi: 10.1186/2047-217X-2-10.
- [97] Wang, W., Wei, Z., Lam, T. et al. Next generation sequencing has lower sequence coverage and poorer SNP-detection capability in the regulatory regions. *Scientific Reports*, (55), 2011. doi: 10.1038/srep00055.
- [98] Mayer, K., Taudien, S., Martis, M. et al. Gene content and virtual gene order of barley chromosome 1H. *Plant Physiology*, 151:496–505, 2009.
- [99] Mayer, K., Martis, M., Hedley, P. et al. Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*, 23(4):1249–63, 2011.
- [100] Hernandez, P., Martis, M., Dorado, G. et al. Next-generation sequencing and syntenic integration of flow-sorted arms of wheat chromosome 4A exposes the chromosome structure and gene content. *Plant Journal*, 69(3):377–86, 2012.
- [101] Brenchley, R., Spannagl, M., Pfeifer, M. et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. *Nature*, 491:705–710, 2012. doi: 10.1038/nature11650.
- [102] Martis, M., Zhou, R., Haseneyer, G. et al. Reticulate evolution of the rye genome. *Plant Cell*, 25:3685–3698, 2013.
- [103] Shantz, H. The place of grasslands in the earth’s cover. *Ecology*, 35(2):143–145, 1954.
- [104] Kellogg, E. Evolution history of the grasses. *Plant Physiology*, 125:1198–1205, 2001.
- [105] Gaut, B. Evolution dynamics of grass genomes. *New Phytologist*, 154:15–28, 2002.

BIBLIOGRAPHY

- [106] Group, G.P.W. Phylogeny and subfamilial classification of the grasses (*Poaceae*). *Annals of the Missouri Botanical Garden*, 88(3):373–457, 2001.
- [107] Hsiao, C., Jacobs, S., Chatterton, N. et al. A molecular phylogeny of the grass family (*Poaceae*) based on the sequences of nuclear ribosomal DNA (ITS). *Australian Systematic Botany*, 11(5-6):667–688, 1999.
- [108] Keller, B. and Feuillet, C. Colinearity and gene density in grass genomes. *Trends in Plant Science*, 2000.
- [109] Matsuoka, Y., Vigouroux, Y., Goodman, M. et al. A single domestication for maize shown by multilocus microsatellite genotyping. *Proc Natl Acad Sci U S A.*, 99:6080–4, 2002.
- [110] Piperno, D. and Flannery, K. The earliest archaeological maize (*Zea mays L.*) from high Mexico: new accelerator mass spectrometry dates and their implications. *Proc Natl Acad Sci U S A.*, 98:2101–3, 2001.
- [111] Smith, B. The initial domestication of *Curcubita pepo* in the Americas 10,000 years ago. *Science*, 276:932–934, 1997.
- [112] Lantin, R. Rice: Post-harvest operations. INPhO - Post-harvest compendium., October 1999. URL http://www.fao.org/fileadmin/user_upload/inpho/docs/Post_Harvest_Compendium_-_RICE.pdf. FAO.
- [113] Zohary, D., Hopf, M. and Weiss, E. *Domestication of plants in the old world: the origin and spread of domesticated plants in southwest Asia, Europe, and the Mediterranean basin*. Oxford Univ. Press, fourth edition, 2012. (first edition published 1988).
- [114] Liu, L., Lee, G., Jiang, L. et al. Evidence for the early beginning (c. 9000 cal. BP) of rice domestication in China: a response. *The Holocene*, 17(8):1059–1068, 2007.
- [115] Fuller, D., Sato, Y., Castillo, C. et al. Consilience of genetics and archaeobotany in the entangled history of rice. *Archaeol Anthropol Sci*, 2:115–131, 2010.
- [116] Molina, J., Sikora, M., Garud, N. et al. Molecular evidence for a single evolution origin of domesticated rice. *PNAS*, 108(20):8351–8356, 2011.
- [117] Huang, X., Kurata, N., Wei, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature*, 490:497–501, 2012.
- [118] Baloch, U. Wheat post-harvest operations. INPhO - Post-harvest compendium., October 1999. URL http://www.fao.org/fileadmin/user_upload/inpho/docs/Post_Harvest_Compendium_-_WHEAT.pdf. FAO.
- [119] Harlan, J. and Zohary, D. Distribution of wild wheat and barley. *Science*, 153(3740): 1074–80, 1966.
- [120] Lev-Yadun, S., Gopher, A. and Abbo, S. The cradle of agriculture. *Science*, 288(5471): 1602–1603, June 2000. doi: 10.1126/science.288.5471.1602.

- [121] Salamini, F., Özkan, H., Brandolini, A. et al. Genetic and geography of wild cereal domestication in the near east. *Nat. Rev. Genet.*, 3:429–441, 2002. pmid: 12042770.
- [122] Heun, M., Schäfer-Pregl, R., Klawan, D. et al. Site of Einkorn wheat domestication identified by DNA fingerprinting. *Science*, 278:1312–1314, 1997.
- [123] Collins, H., Burton, R., Topping, D. et al. Variability in fine structures of noncellulosic cell wall polysaccharides from cereal grains: potential importance in human health and nutrition. *Cereal Chemistry*, 87:272–282, 2010.
- [124] Doležel, J., Greilhuber, J., Lucretti, S. et al. Plant genome size estimation by flow cytometry: inter-laboratory comparison. *Ann. Bot.*, 82:17–26, 1998.
- [125] Allard, R. History of plant population genetics. *Annu. Rev. Genet.*, 33:1–27, 1999. doi: 10.1146/annurev.genet.33.1.1.
- [126] Abbo, S., Lev-Yadun, S. and Gopher, A. Agricultural origins: centers and noncenters; a near eastern reappraisal. *Critical Reviews in Plant Sciences*, 29:317–328, 2010. doi: 10.1080/07352689.2010.502823.
- [127] Badr, A., Müller, K., Schäfer-Pregl, R. et al. On the origin and domestication history of barley (*Hordeum vulgare*). *Mol. Biol. Evol.*, 17(4):499–510, 2000.
- [128] Badr, A. and El-Shazly, H. Molecular approach to origin, ancestry and domestication history of crop plants: barley and clover as examples. *Journal of Genetic Engineering and Biotechnology*, 10:1–12, 2012. doi: 10.1016/j.jgeb.2011.08.002.
- [129] Molina-Cano, J., Moralejo, M., Igartua, E. et al. Further evidence supporting Morocco as a centre of origin of barley. *Theor Appl Genet*, 98:913–918, 1999.
- [130] Ren, X. Tibet as a potential domestication center of cultivated barley of China. *PLoS One*, 8(5):e62700, 2013. doi: 10.1371/journal.pone.0062700.
- [131] Dai, F., Nevo, E., Wu, D. et al. Tibet is one of the centers of domestication of cultivated barley. *PNAS*, 109(42):16969–16973, 2012.
- [132] Dai, F., Chen, Z., Wang, X. et al. Transcriptome profiling reveals mosaic genomic origins of modern cultivated barley. *PNAS*, 111(37):13403–13408, 2014.
- [133] Diamond, J. Evolution, consequences and future of plant and animal domestication. *Nature*, 418:700–707, 2002.
- [134] Hillman, G., Hedges, R., Moore, A. et al. New evidence of lateglacial cereal cultivation at Abu Hureyra on the Euphrates. *The Holocene*, 2001.
- [135] Purugganan, M. and Fuller, D. The nature of selection during plant domestication. *Nature*, 457, 2009. doi: 10.1038/nature07895.

BIBLIOGRAPHY

- [136] Hillman, G. and Davies, M. Measured domestication rates in wild wheat and barley under primitive cultivation, and their archaeological implications. *Journal of World Prehistory*, 4(2), 1990.
- [137] Sencer, H. and Hawkes, J. On the origin of cultivated rye. *Biological Journal of the Linnean Society of London*, 13(4):299–313, 1980.
- [138] Roshevitz, R. A monograph of wild, weedy and cultivated species of rye. *Trudy Botanicheskij Institut Akademija Nauk SSSR*, 6:105–163, 1947.
- [139] Frederiksen, S. and Petersen, G. A taxonomic revision of *Secale* (*Triticeae*, *Poaceae*). *Nordic Journal of Botany*, 18:399–420, 1998.
- [140] Hammer, K., Skolimowska, E. and Knüpffer, L. Vorarbeiten zur monographischen Darstellung von Wildpflanzensortimenten: *Secale* l. *Die Kulturpflanze*, 35(2):135–177, 1987.
- [141] De Bustos, A. and Jouve, N. Phylogenetic relationships of the genus *Secale* based on the characterisation of rDNA ITS sequences. *Plant Syst. Evol.*, 2002.
- [142] Shang, H., Wei, Y., Wang, X. et al. Genetic diversity and phylogenetic relationships in the rye genus *Secale* L. (rye) based on *Secale cereale* microsatellite markers. *Genetics and Molecular Biology*, 29(4):685–691, 2006.
- [143] Chikmawati, T., Skovmand, B. and Gustafson, J. Phylogenetic relationships among *Secale* species revealed by amplified fragment length polymorphisms. *Genome*, 48(5):792–801, 2005.
- [144] Skuza, L., Rogalska, S. and Bocianowski, J. RFLP analysis of mitochondrial DNA in the genus *Secale*. *Acta Biologica Cracoviensia Series Botanica*, 49(1):77–87, 2007.
- [145] Achrem, M., Kalinka, A. and Rogalska, S. Assessment of genetic relationships among *Secale* taxa by using ISSR and IRAP markers and the chromosomal distribution of the AAC microsatellite sequence. *Turkish Journal of Botany*, 38:213–225, 2014.
- [146] Hillman, G. On the origin of domestic rye - *Secale cereale*: the finds from aceramic Can Hasan III in Turkey. *Anatolian Studies*, 28:157–174, 1978.
- [147] Willcox, G. The distribution, natural habitats and availability of wild ccereal in relation to their domestication in the Near East: multiple events, multiple centres. *Veget Hist Archaeobot*, 14:534–541, 2005.
- [148] Behre, K. The history of rye cultivation in Europe. *Veget Hist Archaeobot*, 1:141–156, 1992.
- [149] Searchinger, T., Hanson, C., Ranganathan, J. et al. "The great balancing act". Working paper, Installment 1 of Creating a sustainable food future., May 2013. URL <http://www.worldresourcesreport.org>.

- [150] Bennetzen, J. and Freeling, M. Grasses as a single genetic system: genome composition, collinearity and compatibility. *Trends Genetics*, 9:259–261, 1993. doi: 10.1016/0168-9525(93)90001-X.
- [151] Devos, K. and Gale, M. Genome relationships: the grass model in current research. *The Plant Cell*, 12:637–646, 2000.
- [152] Mehboob-ur-Rahman and Paterson, A. *Molecular techniques in crop improvement. Chapter 2: Comparative genomics in crop plants*. Springer, 2nd edition, 2010. doi: 10.1007/978-90-481-2967-6.
- [153] Kilian, A., Chen, J., Han, F. et al. Towards map-based cloning of the barley stem rust resistance genes *Rpg1* and *rpg4* using rice as an intergenomic cloning vehicle. *Plant Molecular Biology*, 35:187–195, 1997.
- [154] Moore, G., Devos, K., Wang, Z. et al. Grasses, line up and form a circle. *Current Biology*, 5(7):737–739, 1995.
- [155] Devos, K. and Gale, M. Comparative genetics in the grasses. *Plant Molecular Biology*, 35:3–15, 1997.
- [156] Devos, K. Updating the ‘crop circle’. *Current Opinion in Plant Biology*, 8:155–162, 2005. doi: 10.1016/j.pbi.2005.01.005.
- [157] Bennetzen, J. Patterns in grass genome evolution. *Current Opinion in Plant Biology*, 10: 176–181, 2007.
- [158] Kurata, N., Moore, G., Nagamura, Y. et al. Conservation of genome structure between rice and wheat. *Nature Biotechnology*, 12:276–278, 1993.
- [159] Salse, J. *In silico* archeogenomics unveils modern plant genome organisation, regulation and evolution. *Current Opinion in Plant Biology*, 15:122–130, 2012.
- [160] Naranjo, T. and Fernández-Rueda, P. Homoeology of rye chromosome arms to wheat. *Theor Appl Genet*, 82:577–586, 1991.
- [161] Devos, K., Atkinson, M., Chinoy, C. et al. Chromosomal rearrangement in the rye genome relative to that of wheat. *Theor Appl Genet*, 85:673–680, 1993.
- [162] Devos, K., Dubcovsky, J., Devorak, J. et al. Structural evolution of wheat chromosomes 4A, 5A, and 7B and its impact on recombination. *Theor Appl Genet*, 91:282–288, 1995.
- [163] Zhang, H., Jia, J., Gale, M. et al. Relationships between the chromosomes of *Aegilops umbellulata* and wheat. *Theor Appl Genet*, 96:69–75, 1998.
- [164] Zhang, H., Reader, S., Liu, X. et al. Comparative genetic analysis of the *Aegilops longissima* and *Aegilops sharonensis* genomes with common wheat. *Theor Appl Genet*, 103: 518–525, 2001.

BIBLIOGRAPHY

- [165] Salina, E., Leonova, I., Tatyana, T. et al. Wheat genome structure: translocations during the course of polyploidization. *Functional & Integrative Genomics*, 6:71–80, 2006. doi: 10.1007/s10142-005-0001-4.
- [166] Gill, K., Lubbers, E., Gill, B. et al. A genetic linkage map of *Triticum tauschii* (DD) and its relationship to the D genome of bread wheat (AABBDD). *Genome*, 34(3):362–374, 1991.
- [167] Jones, E., Mahoney, N., Hayward, M. et al. An enhanced molecular marker based genetic map of perennial ryegrass (*Lolium perenne*) reveals comparative relationships with other *Poaceae* genomes. *Genome*, 45(2):282–295, 2002.
- [168] Pfeifer, M., Martis, M., Asp, T. et al. The perennial ryegrass genome: targeted use of genome resources for comparative grass genomics. *Plant Physiology*, 161(2):571–82, 2013.
- [169] Alm, V., Fang, C., Busso, C. et al. A linkage map of meadow fescue (*Festuca pratensis* Huds.) and comparative mapping with other *Poaceae* species. *Theor Appl Genet*, 108: 25–40, 2003.
- [170] Kopecký, D., Martis, M., Cíhalíková, J. et al. Flow sorting and sequencing meadow fescue chromosome 4F. *Plant Physiology*, 163(3):1323–37, 2013.
- [171] Sim, S., Chang, T., Curley, J. et al. Chromosomal rearrangement differentiating the ryegrass genome from the *Triticeae*, oat, and rice genomes using common heterologous RFLP probes. *Theor Appl Genet*, 110:1011–1019, 2005.
- [172] Van Deynze, A., Nelson, J., O'Donoghue, L. et al. Comparative mapping in grasses. oat relationships. *Mol Gen Genet*, 249:349–356, 1995.
- [173] Petersen, G., Seberg, O., Yde, M. et al. Phylogenetic relationships of *Triticum* and *Aegilops* and evidence for the origin of the A, B, and D genomes of common wheat (*Triticum aestivum*). *Molecular Phylogenetics and Evolution*, 39:70–82, 2006.
- [174] Chen, M., SanMiguel, P., De Oliveira, A. et al. Microcollinearity in *sh2*-homologous regions of the maize, rice and sorghum genomes. *PNAS*, 94:3431–3435, 1997.
- [175] Tikhanov, A., SanMiguel, P., Nakajima, Y. et al. Colinearity and its exceptions in orthologous *adh* regions of maize and sorghum. *PNAS*, 96:7409–7414, 1999.
- [176] Wicker, T., Stein, N., Albar, L. et al. Analysis of a contiguous 211 kb sequence in diploid wheat (*T. monococcum* l.) reveals multiple mechanisms of genome evolution. *Plant Journal*, 26:307–316, 2001.
- [177] Feuillet, C. and Keller, B. Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *Annals of Botany*, 89:3–10, 2002. doi: 10.1093/aob.2002.mcf008.

- [178] Gottlieb, A., Müller, H., Massa, A. et al. Insular organization of gene space in grass genomes. *PLoS One*, 8(1):e54101, 2013. doi: 10.1371/journal.pone.0054101.
- [179] Sidhu, D. and Gill, K. Distribution of genes and recombination in wheat and other eukaryotes. *Plant Cell, Tissue and Organ Culture*, 4653PB:1–14, 2004.
- [180] Raats, D., Frenkel, Z., Krugman, T. et al. The physical map of wheat chromosome 1BS provides insights into its gene space organisation and evolution. *Genome Biology*, 14:R138, 2013.
- [181] Minx, P., Cordum, H., Wilson, R. et al. Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. *Genome Research*, 15:1284–1291, 2005.
- [182] Schnable, P., Ware, D., Fulton, R. et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*, 326:1112–1115, 2009.
- [183] Bennetzen, J., Schmutz, J., Wang, H. et al. Reference genome sequence of the model plant *Setaria*. *Natur Biotechnology*, 2012.
- [184] Devos, K., Ma, J., Pontaroli, A. et al. Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc. Natl. Acad. Sci. USA*, 102:19243–19248, 2005.
- [185] Matsumoto, T., Tanaka, T., Sakai, H. et al. Comprehensive sequence analysis of 24,783 barley full-length cDNAs derived from twelve clone libraries. *Plant Physiology*, 156:20–28, 2011.
- [186] Wilson, E. The supernumerary chromosomes of *Hemiptera*. *Science*, 26:870–71, 1907.
- [187] Randolph, L. Types of supernumerary chromosomes in maize. *Anatomical Record*, 41:102, 1928.
- [188] Battaglia, E. Cytogenetics of B-chromosomes. *Caryologia*, 17, n. 1, 1964.
- [189] Jones, R. and Rees, H. *B chromosomes*. Academic Press, 1982.
- [190] Jones, R. and Houben, A. B chromosomes in plants: escapees from the A chromosome genome? *Trends in Plant Science*, 8:417–423, 2003.
- [191] Jones, R., Viegas, W. and Houben, A. A century of B chromosomes in plants: So what? *Annals of Botany*, 101:767–775, 2008.
- [192] Jones, R., Gonzales-Sanchez, M., Gonzales-Garcia, M. et al. Chromosomes with a life of their own. *Cytogene*, 120:265–280, 2008.
- [193] Houben, A., Banaei-Moghaddam, A., Klemme, S. et al. Evolution and biology of supernumerary B chromosomes. *Cell. Mol. Life Sci.*, 2014.

BIBLIOGRAPHY

- [194] Mendelson, D. and Zohary, D. Behaviour and transmission of supernumerary chromosomes in *Aegilops speltoides*. *Heredity*, 29:329–339, 1972.
- [195] Gotoh, K. Über die Chromosomenzahl von *Secale cereale* L. *Botanical Magazine Tokyo*, 38:113–122, 1924.
- [196] Kuwada, Y. On the number of chromosomes in maize. *Botanical Magazine Tokyo*, 39:227–234, 1925.
- [197] Longley, A. Supernumerary chromosomes in *Zea mays*. *Journal of Agricultural Research*, 35:769–784, 1927.
- [198] Palestis, B., Trivers, R., Burt, A. et al. The distribution of B chromosomes across species. *Cytogenetic and Genome Research*, 106:151–158, 2004.
- [199] Trivers, R., Burt, A. and Palestis, B. B chromosomes and genome size in flowflow plants. *Genome*, 47:1–8, 2004.
- [200] Palestis, B., Burt, A., Jones, R. et al. B chromosomes are more frequent in mammals with acrocentric karyotypes: support for the theory of centromeric drive. *Proc R Soc Lond B*, 271:S22–S24, 2004.
- [201] Liehr, T., Mrasek, K., Kosyakova, N. et al. Small supernumerary marker chromosomes (sSMC) in humans; are there B chromosomes hidden among them. *Molecular Cytogenetics*, 1(1):12, 2008. ISSN 1755-8166. doi: 10.1186/1755-8166-1-12.
- [202] Houben, A., Leach, C., Verlin, D. et al. A repetitive DNA sequence common to the different B chromosomes of the genus *Brachycome*. *Chromosoma*, 106:513–9, 1997.
- [203] Ziegler, C., Lamatsch, D., Steinlein, C. et al. The giant B chromosome of the cyprinid fish *Alburnus alburnus* harbours a retrotransposon-derived repetitive DNA sequence. *Chromosome Research*, 11:23–35, 2003.
- [204] Camacho, J., Sharbel, T. and Beukeboom, L. B-chromosome evolution. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 355:163–178, 2000.
- [205] Beukeboom, L. Bewildering Bs: an impression of the 1st B-chromosome conference. *Heredity*, 73:328–336, 1994.
- [206] Carter, C. The cytology of *Brachycome*. 8. the inheritance, frequency and distribution of B chromosomes in *B. dichromosomatica* (n=2), formerly in *B. lineariloba*. *Chromosoma*, 67:109–121, 1978.
- [207] Beukeboom, L., Seif, M., Mettenmeyer, T. et al. Paternal inheritance of B chromosomes in a parthenogenetic hermaphrodite. *Heredity*, 77:646–654, 1996.
- [208] Camacho, J., Shaw, M., López-León, M. et al. Population dynamics of a selfish B chromosome neutralized by the standard genome in the grasshopper *Eyprepocnemis plorans*. *Am. Nat.*, 149:1030–1050, 1997.

- [209] Green, D., Zeyl, C. and Sharbel, T. The evolution of hypervariable sex and supernumerary (B) chromosomes in the relict New Zealand frog, *Leiopelma hochstetteri*. *J. evol. Biol.*, 1993.
- [210] Jamilena, M., Ruiz Rejon, C. and Ruiz Rejon, M. A molecular analysis of the origin of the *Crepis capillaris* B chromosome. *Journal of Cell Science*, 107:703–708, 1994.
- [211] Dhar, M., Friebe, B., Koul, A. et al. Origin of an apparent B chromosome by mutation, chromosome fragmfragment and specific DNA sequence amplification. *Chromosoma*, 111:332–340, 2002.
- [212] Sapre, A. and Deshpande, D. Origin of B chromosomes in *Coix L.* through spontaneous interspecific hybridization. *The Journal of Heredity*, 78:191–196, 1987.
- [213] McAllister, B. and Werren, J. Hybrid origin of a B chromosome (PSR) in the parasitic wasp *Nasonia vitripennis*. *Chromosoma*, 106:243–253, 1997.
- [214] Berdnikov, V., Gorel, F., Kosterin, O. et al. Tertiary trisomics in the garden pea as a model of B chromosome evolution in plants. *Heredity*, 91:577–583, 2003.
- [215] Stark, E., Connerton, I., Bennett, S. et al. Molecular analysis of the structure of the maize B-chromosome. *Chromosome Research*, 4:15–23, 1996.
- [216] Houben, A., Kynast, R., Heim, U. et al. Molecular cytogenetic characterisation of the terminal heterochromatic segment of the B-chromosome of rye (*Secale cereale*). *Chromosoma*, 105:97–103, 1996.
- [217] López-León, M., Neves, N., Schwarzacher, T. et al. Possible origin of a B chromosome deduced from its DNA composition using double FISH technique. *Chromosome*, 1994.
- [218] Green, D. Cytogenetics of the endemic New Zealand frog, *Leiopelma hochstetteri*: extraordinary supernumerary chromosome variation and a unique sex-chromosome system. *Chromosoma*, 97:55–70, 1988.
- [219] Perfectti, F. and Werren, J. The interspecific origin of B chromosomes: experimental evidence. *Evolution*, 55(5):1069–1073, 2001.
- [220] Porto, F., Portela-Castro, A. and Martins-Santos, I. Possible origins of B chromosomes in *Rineloricaria pentamaculata* (Loricariidae, Siluriformes) from the Parana River basin. *Genet. Mol. Res*, 9(3):1654–1659, 2010.
- [221] Houben, A., Verlin, D., Leach, C. et al. The genomic complex of micro B chromosomes of *Brachycome dichromosomatica*. *Chromosoma*, 110:451–459, 2001.
- [222] Sharbel, T., Green, D. and Houben, A. B-chromosome origin in the endemic New Zealand frog *Leiopelma hochstetteri* through sex chromosome devolution. *Genome*, 41:14–22, 1998.

BIBLIOGRAPHY

- [223] Yoshida, K., Terai, Y., Mizoiri, S. et al. B chromosomes have a functional effect on female sex determination in lake Victoria cichlid fishes. *PLOS Genetics*, 7, 2011.
- [224] Schartl, M., Nanda, I., Schlupp, I. et al. Incorporation of subgenomic amounts of DNA as compensation for mutational load in a gynogenetic fish. *Nature*, 373:68–71, 1995.
- [225] Jackson, R. and Newmark, P. Effect of supernumerary chromosomes on production of pigment in *Haplopappus gracilis*. *Science*, 132:1316–1317, 1960.
- [226] Randolph, L. Genetic characteristics of the B chromosomes in maize. *Genetics*, 26: 608–631, 1941.
- [227] Dherawat, A. and Sadanaga, K. Cytogenetics of a crown rust-resistant hexaploid oat with 42 + 2 fragment chromosomes. *Crop Sci*, 13(6):591–594, 1973.
- [228] Kousaka, R. and Endo, T. Effect of a rye B chromosome and its segments on homoeologous pairing in hybrids between common wheat and *Aegilops variabilis*. *Genes Genet. Syst.*, 87:1–7, 2012.
- [229] Staub, R. Leaf striping correlate with the presence of B chromosomes in maize. *J Hered*, 78(2):71–74, 1987.
- [230] Beukeboom, L., Seif, M., Plowman, A. et al. Phenotypic fitness effects of B chromosomes in the pseudogamous parthenogenetic planarian *Polycelis nigra*. *Heredity*, 80:594–603, 1998.
- [231] Procnier, W. B chromosomes of *Cnephia dacotensis* and *C. ornithophilia* (Diptera: Simuliidae). *Can. J. Zool.*, 53:1638–1647, 1975.
- [232] Meratan, A., Ghaffari, S., Niknam, V. et al. Antioxidative responses in calli of two populations of *Acanthophyllum laxiusculum* with and without B-chromosomes under salt stress. *Pakistan Journal of Biological Sciences*, 16(1), 2013.
- [233] Holmes, D. and Bougourd, S. B-chromosome selection in *Allium schoenoprasum*. i. natural populations. *Heredity*, 63:83–87, 1989.
- [234] Holmes, D. and Bougourd, S. B-chromosome selection in *Allium schoenoprasum* ii. experimental populations. *Heredity*, 67:117–122, 1991.
- [235] Plowman, A. and Bougourd, S. Selectively advantageous effects of B chromosomes on germination behaviour in *Allium schoenoprasum* L. *Heredity*, 72:587–593, 1994.
- [236] Teoh, S., Rees, H. and Hutchinson, J. B chromosome selection in *Lolium*. *Heredity*, 37 (2):207–213, 1976.
- [237] Teoh, S. and Jones, R. B chromosome selection and fitness in rye. *Heredity*, 41(1):35–48, 1978.
- [238] Hutchinson, J. Selection of B chromosomes in *Secale cereale* and *Lolium perenne*. *Heredity*, 34:39–52, 1975.

- [239] Han, Y., Liu, X., Benny, U. et al. Genes determining pathogenicity to pea are clustered on a supernumerary chromosome in the fungal plant pathogen *Nectria haematococca*. *The Plant Journal*, 25(3):305–314, 2001.
- [240] Houben, A., Belyaev, N., Leach, C. et al. Differences of histone H4 acetylation and replication timing between A and B chromosomes of *Brachycome dichromosomatica*. *Chromosome Research*, 5:233–237, 1997.
- [241] Marschner, S. B chromosomes of *B. dichromosomatica* show a reduced level of euchromatic histone H3 methylation marks. *Chromosome Research*, 15:215–222, 2007.
- [242] Schmid, M., Ziegler, C., Steinlein, C. et al. Chromosome banding in *Amphibia*. xxiv. the B chromosomes of *Gastrotheca espeletia* (Anura, Hylidae). *Cytogenetic Genome Res*, 97(3-4):205–18, 2002.
- [243] Raman, R. and Sharma, T. Dna replication, G- and C-bands and meiotic behaviour of supernumerary chromosomes of *Rattus rattus* (Linn.). *Chromosoma*, 45:111–119, 1974.
- [244] Świtoński, M., Gustavsson, I., Höjer, K. et al. Synaptonemal complex analysis of the B-chromosomes in spermatocytes of the silver fox (*vulpes fulvus* desm.). *Cytogenetic Cell Genet*, 45:84–92, 1987.
- [245] Maistro 1992. Occurrence of macro B chromosomes in *Astyanax scabripinnis* paranae (Pisces, Characiformes, Characidae). *Genetica*, 87:101–106, 1992.
- [246] Timmis, J. N., Ingle, J. and Sinclair, J. The genomic quality of rye B chromosomes. *Journal of Experimental Botany*, 26(92):367–378, 1975.
- [247] Rimpau, J. and Flavell, R. Characterisation of rye B chromosome DNA by DNA/DNA hybridisation. *Chromosoma*, 52:207–217, 1975.
- [248] Wilkes, T., Francki, M., Langridge, P. et al. Analysis of rye B-chromosome structure using fluorescence in situ hybridization (FISH). *Chromosome Research*, 3:466–472, 1995.
- [249] Sandery, M., Forster, J., Blunden, R. et al. Identification of a family of repeated sequences on the rye B chromosome. *Genome*, 33(6):908–913, 1990.
- [250] Houben, A., Field, B. and Saunders, V. Microdissection and chromosome painting of plant B chromosomes. *Methods Cell Sci*, 23(1-3):115–124, 2001.
- [251] Kubaláková, M., Valárik, M., Bartoš, J. et al. Analysis and sorting of rye (*Secale cereale* L.) chromosomes using flow cytometry. *Genome*, 2003.
- [252] Lamb, J., Riddle, N., Cheng, Y. et al. Localization and transcription of a retrotransposon-derived element on the maize B chromosome. *Chromosome Research*, 2007.
- [253] Langdon, T., Seago, C., Jones, R. et al. De novo evolution of satellite DNA on the rye B chromosome. *Genetics*, 2000.

BIBLIOGRAPHY

- [254] Green, D. Muller's ratchet and the evolution of supernumerary chromosomes. *Genome*, 33:818–824., 1990.
- [255] Gutknecht, J., Sperlich, D. and Bachmann, L. A species specific satellite DNA family of *Drosophila subsilvestris* appearing predominantly in B chromosomes. *Chromosoma*, 103:539–544, 1995.
- [256] Alfenito, M. and Birchler, J. Molecular characterisation of a maize B chromosome centric sequence. *Genetics*, 135:589–597, 1993.
- [257] Blunden, R., Wilkes, T., Forster, J. et al. Identification of the E3900 family, a second family of rye B chromosome specific repeated sequences. *Genome*, 36(4):706–11, 1993.
- [258] Cheng, Y. and Lin, B. Molecular organisation of large fragment in the maize B chromosome: indication of a novel repeat. *Genetics*, 166:1947–1961, 2003.
- [259] Beukeboom, L. and Werren, J. Deletion analysis of the selfish B chromosome, paternal sex ratio (PSR), in the parasitic wasp *Nasonia vitripennis*. *Genetics*, 133:637–648, 1993.
- [260] McQuade, L., Hill, R. and Francis, D. B-chromosome system in the greater glider, *Petauroides volans* (Marsupialia: Pseudocheiridae). ii. investigation of B-chromosome DNA sequences isolated by micromanipulation and PCR. *Cytogenetics and Cell Genetics*, 66(3):155–61, 1994.
- [261] Leach, C., Donald, T., Franks, T. et al. Organisation and origin of a B chromosome centromeric sequence from *Brachycome dichromosomatica*. *Chromosoma*, 103:708–714, 1995.
- [262] Carchilan, M., Delgado, M., Ribeiro, T. et al. Transcriptionally active heterochromatin in rye B chromosomes. *Plant Cell*, 19:1738–1749, 2007.
- [263] John, U., Leach, C. and Timmis, J. A sequence specific to B chromosomes of *Brachycome dichromosomatica*. *Genome*, 34(5):739–744, 1991.
- [264] Nur, U., Werren, J., Eickbush, D. et al. A 'selfish' B chromosome that enhances its transmission by eliminating the paternal genome. *Science*, 240(4851):512–514, 1988.
- [265] Trifonov, V., Perelman, P., Kawada, S. et al. Complex structure of B-chromosomes in two mammalian species: *Apodemus peninsulae* (Rodentia) and *Nyctereutes procyonoides* (Carnivora). *Chromosome Res*, 10(2):109–116, 2002.
- [266] Pedro, J. and Camacho, M. *B chromosomes*. In: Gregory, T.R. ed. *The evolution of the genome*: Elsevier/Academic Press, 2005.
- [267] Puertas, M. Nature and evolution of B chromosomes in plants: a non-coding but information-rich part of plant genomes. *Cytogenetic and Genome Research*, 2002.
- [268] Tanić, N., Vujošević, M., Dedović-Tanić, N. et al. Differential gene expression in yellow-necked mice *Apodemus flavicollis* (Rodentia Mammalia) with B chromosomes. *Chromosoma*, 113:418–427, 2005.

- [269] Graphodatsky, A., Kukekova, A., Yudkin, D. et al. The proto-oncogene C-KIT maps to canid B-chromosome. *Chromosome Research*, 13:113–122, 2005.
- [270] Donald, T., Leach, C., Clough, A. et al. Ribosomal RNA genes and the B chromosome of *Brachycome dichromosomatica*. *Heredity*, 74:556–561, 1995.
- [271] Poletto, A., Ferreira, I. and Martins, C. The B chromosomes of the African cichlid fish *Haplochromis obliquoidens* harbour 18S rRNA gene copies. *BMC Genetics*, 2010.
- [272] Marschner, S., Meister, A., Blattner, F. et al. Evolution and function of B chromosome 45S rDNA sequences in *Brachycome dichromosomatica*. *Genome*, 50(7):638–644, 2007.
- [273] Cabrero, J., Alché, J. and Camacho, J. Effect of B chromosomes on the activity of nuclear organizer regions in the grasshopper *Eyprepocnemis plorans*: activation of a latent nucleolar organizer region on a B chromosome fused to an autosome. *Genome*, 29(1): 116–121, 1987.
- [274] López-León, M., Cabrero, J. and Camacho, J. A nucleolus organiser region in a B chromosome inactivated by DNA methylation. *Chromosoma*, 100:134–138, 1991.
- [275] Maluszynska, J. and Schweizer, D. Ribosomal RNA genes in B chromosomes of *Crepis capillaris* detected by non-radioactive in situ hybridisation. *Heredity*, 62:59–65, 1989.
- [276] Leach, C., Houben, A., Field, B. et al. Molecular evidence for transcription of genes on a B chromosome in *Crepis capillaris*. *Genetics*, 171:269–278, 2005.
- [277] Camacho, J., Navas-Castillo, J. and Cabrero, J. Extra nucleolar activity associated with presence of a supernumerary chromosome segment in the grasshopper *Oedipoda fuscocincta*. *Heredity*, 56:237–241, 1986.
- [278] Brockhouse, C., Bass, J., Feraday, R. et al. Supernumerary chromosome evolution in the *Simulium venum* group (Diptera: Simuliidae). *Genome*, 32:516–521, 1989.
- [279] Teruel, M., Cabrero, J., Perfectti, F. et al. B chromosome ancestry revealed by histone genes in the migratory locust. *Chromosoma*, 119:217–225, 2010.
- [280] Oliveira, N., Cabral-de Mello, D., Rocha, M. et al. Chromosomal mapping of rDNAs and H3 histone sequences in the grasshopper *rhammatocerus brasiliensis* (acrididae, gomphocerinae): extensive chromosomal dispersion and co-localization of 5S rDNA/H3 histone clusters in the A complement and B chromosome. *Molecular Cytogenetics*, 4:24, 2011.
- [281] Venere, P., Miyazawa, C. and Galetti, P. New cases of supernumerary chromosomes in characiform fishes. *Genetics and Molecular Biology*, 22(3):345–349, 1999.
- [282] Evans, H. Supernumerary chromosomes in wild populations of the snail *Helix pomatia* L. *Heredity*, 15:129–138, 1960.
- [283] Greilhuber, J. and Speta, F. C-banded karyotypes in the *Scilla hohenackeri* group, *S. persica*, and *Puschkinia* (Liliaceae). *Plant Syst. Evol.*, 126:149–188, 1976.

BIBLIOGRAPHY

- [284] Teruel, M., Cabrero, J., Perfectti, F. et al. Microdissection and chromosome painting of X and B chromosomes in the grasshopper *Eyprepocnemis plorans*. *Cytogenetics and Genome Research*, 125:286–291, 2009.
- [285] Östergren, G. Heterochromatic B-chromosomes in *Anthoxanthum*. *Hereditas*, 33:261–296, 1947.
- [286] Jones, R. B-chromosome drive. *The American Naturalist*, 137:430–442, 1991.
- [287] Nur, U. Mitotic instability leading to an accumulation of B chromosomes in grasshopper. *Chromosoma*, 27:1–19, 1969.
- [288] Roman, H. Mitotic nondisjunction in the case of interchanges involving the B-type chromosome in maize. *Genetics*, 32:391–409, 1947.
- [289] Roman, H. Directed fertilisation in maize. *PNAS*, 34:450–454, 1948.
- [290] Carlson, W. The B chromosome of corn. *Annu. Rev. Genet.*, 12:5–23, 1978.
- [291] Lin, B. Regional control of nondisjunction of the B chromosome in maize. *Genetics*, 90:613–627, 1978.
- [292] González-Sánchez, M., González-Sánchez, E., Molina, E. et al. One gene determines maize B chromosome accumulation by preferential fertilisation; another gene(s) determines their meiotic loss. *Heredity*, 90:122–129, 2003.
- [293] López-León, M., Cabrero, J. and Camacho, J. Negatively assorted gamete fertilization for supernumerary heterochromatin in two grasshopper species. *Heredity*, 76:651–657, 1996.
- [294] Werren, J. The paternal-sex-ratio chromosome of *Nasonia*. *Am. Nat.*, 137:392–402, 1991.
- [295] Martis, M., Klemme, S., Banaei-Moghaddam, A. et al. Selfish supernumerary chromosome reveals its origin as a mosaic of host genome and organellar sequences. *Proc Natl Acad Sci U S A.*, 109(33):13343–6, 2012.
- [296] Romera, F., Vega, J., Diez, M. et al. B chromosome polymorphism in Korean rye populations. *Heredity*, 62:117–21, 1989.
- [297] Jiménez, M., Romera, F., Puertas, M. et al. B-chromosomes in inbred lines of rye (*Secale cereale L.*). I. vigour and fertility. *Genetica*, 92:149–154, 1994.
- [298] Kishikawa, H. Cytogenetic studies of B-chromosomes in rye, *Secale cereale L.*, in Japan. *Agric. Bull. Saga Univ.*, 21:1–81, 1965.
- [299] Marques, A., Banaei-Moghaddam, A., Klemme, S. et al. B chromosomes of rye are highly conserved and accompanied the development of early agriculture. *Annals of Botany*, 2013.

- [300] Niwa, K. and Sakamoto, S. Origin of B chromosomes in cultivated rye. *Genome*, 38(2): 307–312, 1995.
- [301] Jones, R. and Puertas, M. *The B chromosomes of rye (Secale cereale L.)*. In Dhir KK, Saareen TS, eds. *Frontiers in Plant Science Research*. (Dehli (India): Bhagwati Enterprises), pp. 81-112, 1993.
- [302] Müntzing, A. Some main results from investigations of accessory chromosomes. *Hereditas*, 57:432–438, 1967.
- [303] González-Sánchez, M., Chiavarino, A., Jiménez, G. et al. The parasitic effects of rye B chromosomes might be beneficial in the long term. *Cytogenet Genome Res*, 106:386–393, 2004.
- [304] Hasegawa, N. A cytological study on B-chromosome rye. *Cytologia*, 6:68–77, 1934.
- [305] Müntzing, A. Cytological studies of extra fragment chromosomes in rye. v. a new fragment type arisen by deletion. *Hereditas*, 34:435–442, 1948.
- [306] Endo, T., Nasuda, S., Jones, N. et al. Dissection of rye B chromosomes, and nondisjunction properties of the dissected segments in a common wheat background. *Genes & Genetic Systems*, 83(1):23–30, 2008.
- [307] Lindström, J. Transfer to wheat of accessory chromosomes from rye. *Hereditas*, 54: 149–155, 1965.
- [308] Puertas, M., Romera, F. and Delapena, A. Comparison of B chromosome effects on *Secale cereale* and *Secale vavilovii*. *Heredity*, 55:229–234, 1985.
- [309] Gale, M. and Devos, K. Comparative genetic in the grasses. *Proc Natl Acad Sci U S A.*, 95:1971–1974, 1998.
- [310] Philippe, R., Paux, E., Bertin, I. et al. A high density physical map of chromosome 1BL supports evolution studies, map-based cloning and sequencing in wheat. *Genome Biology*, 14(6):R64, 2013.
- [311] Luo, M., Gu, Y., You, F. 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):7940–5, 2013.
- [312] Evans, C. YAML draft 0.1. Yahoo! Tech groups: sml-dev, May 2001. URL <http://www.yaml.org>.
- [313] Spannagl, M., Martis, M., Pfeifer, M. et al. Analysing complex *Triticeae* genomes - concepts and strategies. *Plant Methods*, 2013.
- [314] Nussbaumer, T., Martis, M., Roessner, S. et al. MIPS PlantsDB: a database framework for comparative plant genome research. *Nucleic Acids Research*, 41 (Database issue): D1144–51, 2013.

BIBLIOGRAPHY

- [315] Kurtz, S. The vmatch large scale sequence analysis software. URL <http://www.vmatch.de/>.
- [316] IBGSC, Mayer, K., Waugh, R. et al. A physical, genetic and functional sequence assembly of the barley genome. *Nature*, 491:711–716, 2012.
- [317] Schulte, D., Close, T., Graner, A. et al. The international barley sequencing consortium - at the threshold of efficient access to the barley genome. *Plant Physiology*, 149:142–147, 2009.
- [318] Close, T., Bhat, P., Lonardi, S. et al. Development and implementation of high-throughput SNP genotyping in barley. *BMC Genomics*, 10:582, 2009.
- [319] Bolot, S., Abrouk, M., Masood-Quraishi, U. et al. The ‘inner circle’ of the cereal genomes. *Current Opinion in Plant Biology*, 12:119–125, 2009.
- [320] Waugh, R., Jannink, J., Muehlbauer, G. et al. The emergence of whole genome association scans in barley. *Current Opinion in Plant Biology*, 12:218–222, 2009.
- [321] Helguera, M., Rivarola, M., Clavijo, B. et al. New insights into the wheat chromosome 4D structure and virtual gene order, revealed by survey pyrosequencing. *Plant Science*, 2014.
- [322] Abrouk, M., Klocová, B., Šimková, H. et al. In-silico identification and characterization of wheat 4AL - *Triticum militinae* introgression. manuscript, 2014.
- [323] Abrouk, M., Martis, M., Molnár, I. et al. First large-scale insights into the chromosome structure of two wild relatives of wheat. manuscript, 2014.
- [324] The International Wheat Genome Sequencing Consortium. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*, 345(6194), 2014. doi: 10.1126/science.1251788.
- [325] Poursarebani, N., Ariyadasa, R., Zhou, R. et al. Conserved synteny-based anchoring of the barley genome physical map. *Functional & Integrative Genomics*, 13:339–350, 2013. doi: 10.1007/s10142-013-0327-2.
- [326] Lüpken, T., Stein, N., Perovic, D. et al. Genomics-based high-resolution mapping of the BaMMV/BaYMV resistance gene *rym11* in barley (*Hordeum vulgare* L.). *Theor Appl Genet*, 126:1201–1212, 2013. doi: 10.1007/s00122-013-2047-3.
- [327] Lüpken, T., Stein, N., Perovic, D. et al. High-resolution mapping of the barley Ryd3 locus controlling tolerance to BYDV. *Mol Breeding*, 33:477–488, 2013. doi: 10.1007/s11032-013-9966-1.
- [328] Shahinnia, F., Druka, A., Franckowiak, J. et al. High resolution mapping of dense spike-ar (*dsp.ar*) to the genetic centromere of barley chromosome 7H. *Theor Appl Genet*, 124: 373–384, 2012. doi: 10.1007/s00122-011-1712-7.

- [329] Iehisa, J., Shimizu, A., Sato, K. et al. Discovery of high-confidence single nucleotide polymorphisms from large-scale de novo analysis of leaf transcripts of *Aegilops tauschii*, a wild wheat progenitor. *DNA Research*, 19:487–497, 2012.
- [330] Silvar, C., Perovic, D., Nussbaumer, T. 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. doi: 10.1371/journal.pone.0067336.
- [331] Mizuno, N., Nitta, M., Sato, K. et al. A wheat homologue of *PHYTOCLOCK 1* is a candidate gene conferring the early heading phenotype to einkorn wheat. *Genes & Genetic Systems*, 87(6):357–367, 2012.
- [332] Lucas, S., Akpinar, B., Kantar, M. et al. Physical mapping integrated with syntenic analysis to characterize the gene space of the long arm of wheat chromosome 1A. *PLoS One*, 8:e59542, 2013. doi: 10.1371/journal.pone.0059542.
- [333] Rodríguez-Suárez, C. and Atienza, S. *Hordeum chilense* genome, a useful tool to investigate the endosperm yellow pigment content in the *Triticeae*. *BMC Plant Biology*, 12:200, 2012.
- [334] Sorrells, M., La Rota, M., Bermudez-Kandianis, C. et al. Comparative DNA sequence analysis of wheat and rice genomes. *Genome Research*, 13:1818–1827, 2003.
- [335] Tarchini, R., Biddle, P., Wineland, R. et al. The complete sequence of 340 kb of DNA around the rice *Adh1-Adh2* region reveals interrupted colinearity with maize chromosome 4. *The Plant Cell*, 12(3):381–391, 2000.
- [336] Feuillet, C., Stein, N., Rossini, L. et al. Integrating cereal genomics to support innovation in the *Triticeae*. *Functional & Integrative Genomics*, 12:573–583, 2012. doi: 10.1007/s10142-012-0300-5.
- [337] Mascher, M., Muehlbauer, G., Rokhsar, D. et al. Anchoring and order NGS contig assemblies by population sequencing (POPSEQ). *The Plant Journal*, 76(4):718–727, 2013.
- [338] Doležel, J., Kubaláková, M., Paux, E. et al. Chromosome-based genomics in the cereals. *Chromosome Research*, 15(1):51–66, 2007.
- [339] Akhunov, E., Sehgal, S., Liang, H. et al. Comparative analysis of syntenic genes in grass genomes reveals accelerated rates of gene structure and coding sequence evolution in polyploid wheat. *Plant Physiology*, 161:252–265, 2013.
- [340] Vitulo, N., Albiero, A., Forcato, C. et al. First survey of the wheat chromosome 5A composition through a next generation sequencing approach. *PLoS One*, 6(10):e26421, 2011. doi: 10.1371/journal.pone.0026421.
- [341] Shatalina, M., Wicker, T., Buchmann, J. et al. Genotype-specific SNP map based on whole chromosome 3B sequence information from wheat cultivars Arina and Forno. *Plant Biotechnology Journal*, 11:23–32, 2013.

BIBLIOGRAPHY

- [342] Berkman, P., Visendi, P., Lee, H. et al. Dispersion and domestication shaped the genome of barley wheat. *Plant Biotechnology Journal*, 11:564–571, 2013.
- [343] Alnemer, L., Seetan, R., Bassi, F. et al. Wheat Zapper: a flexible online tool for colinearity studies in grass genomes. *Functional Integr Genomics*, 13:11–17, 2013. doi: 10.1007/s10142-013-0317-4.
- [344] Bennetzen, J. and Freeling, M. The unified grass genome: synergy in synteny. *Genome Research*, 7:301–306, 1997.
- [345] Murat, F., Xu, J., Tannier, E. et al. Ancestral grass karyotype reconstruction unravels new mechanisms of genome shuffling as a resource of plant evolution. *Genome Research*, 20:1545–1557, 2010.
- [346] Qi, L., Echaliier, B., Chao, S. 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*, 168:701–712, 2004.
- [347] Stein, N., Prasad, M., Scholz, U. et al. A 1000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. *Theor Appl Genet*, 114:823–839, 2007.
- [348] Choulet, F., Wicker, T., Rustenholz, C. et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, 22:1686–1701, 2010.
- [349] Berkman, P., Skarshewski, A., Manoli, S. et al. Sequencing wheat chromosome arm 7BS delimits the 7BS/4AL translocation and reveals homoeologous gene conservation. *Theor Appl Genet*, 124:423–432, 2012.
- [350] Van Deynze, A., Dubcovsky, J., Gill, K. et al. Molecular-genetic maps for group 1 chromosomes of *Triticeae* species and their relation to chromosomes in rice and oat. *Genome*, 38(1):45–49, 1995. doi: 10.1139/g95-006.
- [351] Salse, J., Abrouk, M., Bolot, S. et al. Reconstruction of monocotyledonous protochromosomes reveals faster evolution in plants than in animals. *Proc Natl Acad Sci U S A.*, 106:14908–14913, 2009.
- [352] Thiel, T., Graner, A., Waugh, R. et al. Evidence and evolutionary analysis of ancient whole-genome duplication in barley predating the divergence from rice. *BMC Evolutionary Biology*, 9:209, 2009. doi: 10.1186/1471-2148-9-209.
- [353] Qi, L., Friebe, B. and Gill, B. Complex genome rearrangement reveal evolutionary dynamics of pericentromeric regions in the *Triticeae*. *Genome*, 49:1628–1639, 2006.
- [354] Hohmann, U., Graner, A., Endo, T. et al. Comparison of wheat physical maps with barley linkage maps for group 7 chromosomes. *Theor Appl Genet*, 91:618–626, 1995.

- [355] Linde-Laursen, I., Heslop-Harrison, J., Shepherd, K. et al. The barley genome and its relationship with the wheat genomes. a survey with an international agreed recommendation for barley chromosome nomenclature. *Hereditas*, 126(1):1–16, 1997.
- [356] Dubcovsky, J., Luo, M., Zhong, G. et al. Genetic map of diploid wheat, *Triticum monococcum* l., and its comparison with maps of *Hordeum vulgare* l. *Genetics*, 143:983–999, 1996.
- [357] Naranjo, T., Roca, A., Goicoechea, P. et al. Arm homoeology of wheat and rye chromosomes. *Genome*, 29:873–882, 1987.
- [358] Miftahudin, Ross, K., Ma, X. et al. Analysis of expressed sequence tag loci on wheat chromosome group 4. *Genetics*, 168:651–663, 2004.
- [359] Ma, J., Stiller, J., Wei, Y. et al. Extensive pericentric rearrangement in the bread wheat (*Triticum aestivum* l.) genotype “Chinese Spring” revealed from chromosome shotgun sequence data. *Genome Biol. Evol.*, 6(11):3039–3048, 2014. doi: 10.1093/gbe/evu237.
- [360] Escobar, J., Scornavacca, C., Cenci, A. et al. Multigenic phylogeny and analysis of tree incongruences in *Triticeae* (*Poaceae*). *BMC Evol. Biol.*, 11:181, 2011.
- [361] Kausserud, H. and Schumacher, T. Ribosomal dna variation, recombination and inheritance in the basidiomycete *Trichaptum abietinum*: implications for reticulate evolution. *Heredity*, 91:163–172, 2003.
- [362] Pyle, R. and Randall, J. A review of hybridization in marine angelfishes (*Perciformes: Pomacanthidae*) . *Environ Biol Fishes*, 41:127–145, 1994.
- [363] Schelly, R., Salzburger, W., Koblmüller, S. et al. Phylogenetic relationships of the lamprologine cichlid genus *Lepidiolamprologus* (*Teleostei: Perciformes*) based on mitochondrial and nuclear sequences, suggesting introgressive hybridization. *Mol Phyl Evol*, 38: 426–438, 2006.
- [364] Gilg, M. and Hilbish, T. Patterns of larval dispersal and their effect on the maintenance of a blue mussel hybrid zone in Southwest England. *Evolution*, 57:1061–1077, 2003.
- [365] Hatta, M., Fukami, H., Wang, W. et al. Reproductive and genetic evidence for a reticulate evolutionary history of mass-spawning corals. *Biol Evol*, 16:1607–1613, 1999.
- [366] Prada, C., Schizas, N. and Yoshioka, P. Phenotypic plasticity or speciation? a case from a clonal marine organism. *BMC Evolutionary Biology*, 8:47, 2008. doi: 10.1186/1471-2148-8-47.
- [367] Doyle, J., Doyle, J., Rauscher, J. et al. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (*Glycine* subgenus *Glycine*). *New Phytologist*, 161:121–132, 2003.
- [368] Snowdon, R. Cytogenetics and genome analysis in *Brassica* crops. *Chromosome Res.*, 15:85–95, 2007.

BIBLIOGRAPHY

- [369] Kellogg, E., Appels, R. and Mason-Gamer, R. When gene trees tell different stories: The diploid genera of *Triticeae*. *Syst. Bot.*, 21:312–347, 1996.
- [370] Rieseberg, L., Carter, R. and Zona, S. Molecular tests of the hypothesized hybrid origin of two diploid *Helianthus* species (*Asteraceae*). *Evolution*, 44:1498–1511, 1990.
- [371] Rieseberg, L. Homoploid reticulate evolution in *Helianthus* (*Asteraceae*): Evidence from ribosomal genes. *American Journal of Botany*, 78(9):1218–1237, 1991.
- [372] Kellogg, E. and Bennetzen, J. The evolution of nuclear genome structure in seed plants. *Am. J. Bot.*, 91:1709–1725, 2004.
- [373] Mahelka, V., Kopecký, D. and Paštová, L. On the genome constitution and evolution of intermediate wheatgrass (*Thinopyrum intermedium*: *Poaceae*, *Triticeae*). *BMC Evol. Biol.*, 11:127, 2011.
- [374] Mallet, J. Hybridization as an invasion of the genome. *Trends Ecol. Evol. (Amst.)*, 20: 229–237, 2005.
- [375] Mason-Gamer, R., Burns, M. and Naum, M. Reticulate evolutionary history of a complex group of grasses: Phylogeny of *Elymus* StStHH allotetraploids based on three nuclear genes. *PLoS ONE*, 5:e10989, 2010.
- [376] Linder, C. and Rieseberg, L. Reconstructing patterns of reticulate evolution in plants. *American Journal of Botany*, 91(10):1700–1708, 2004.
- [377] Choler, P., Erschbamer, B., Tribsch, A. et al. Genetic introgression as a potential to widen a species' niche: Insights from alpine *Carex curvula*. *Proc. Natl. Acad. Sci. USA*, 101: 171–176, 2004.
- [378] Rieseberg, L., Raymond, O., Rosenthal, D. et al. Major ecological transitions in wild sunflowers facilitated by hybridisation. *Science*, 301:1211–1216, 2003.
- [379] Levy, A. and Feldman, M. The impact of polyploidy on grass genome evolution. *Plant Physiol*, 130:1587–1593, 2002.
- [380] Salomon, S. and Puchta, H. Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *EMBO J*, 17:6086–6095, 1998.
- [381] Valente, G., Conte, M., Fantinatti, B. et al. Origin and evolution of B chromosomes in the cichlid fish *Astatotilapia latifasciata* based on integrated genomic analyses. *Mol. Biol. Evol.*, 31:2061–2072, 2014.
- [382] Silva, D., Pansonato-Alves, J., Utsunomia, R. et al. Delimiting the origin of a B chromosome by FISH mapping, chromosome painting and DNA sequence analysis in *Astyanax paranae* (Teleostei, Characiformes). *PLoS One*, 9:e94896, 2014.
- [383] Banaei-Moghaddam, A., Martis, M., Macas, J. et al. Genes on B chromosomes: old questions revisited with new tools. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1849(1):64 – 70, 2015.

- [384] Klemme, S., Banaei-Moghaddam, A., Macas, J. et al. High-copy sequences reveal distinct evolution of the rye B chromosome. *New Phytol.*, 199:550–558, 2013.
- [385] Banaei-Moghaddam, A., Meier, K., Karimi-Ashtiyani, R. et al. Formation and expression of pseudogenes on the B chromosome of rye. *Plant Cell*, 25:2536–2544, 2013.
- [386] Trifonov, V., Dementyeva, P., Larkin, D. et al. Transcription of a protein-coding gene on B chromosomes of the Siberian roe deer (*Capreolus pygargus*). *BMC Biology*, 11, 2013.
- [387] Zhou, Q., Zhu, H., Huang, Q. et al. Deciphering neo-sex and B chromosome evolution by the draft genome of *Drosophila albomicans*. *BMC Genomics*, 13:109, 2012.
- [388] Ruban, A., Fuchs, J., Marques, A. et al. B chromosomes of *Aegilops speltoides* are enriched in organelle genome-derived sequences. *PLoS One*, 9(2):e90214, 2014.
- [389] Houben, A. and Schubert, I. Engineered plant minichromosomes: a resurrection of B chromosomes? *The Plant Cell*, 19:2323–2327, 2007.

Curriculum Vitae

Persönliche Daten

Name: Mihaela-Maria Martis
Geburtsdatum und -ort: 17.11.1979 in Sibiu, Rumänien
Familienstand: ledig
Adresse: Jungwirthstrasse 3
80802 München, Deutschland
Telefon: +49 (0) 176 200 724 74
E-Mail: mihaela_martis@web.de

Schul- und Berufsausbildung

- 2012 – heute **externe Doktorandin**
Technische Universität München
Lehrstuhl für Genomorientierte Bioinformatik
Titel: *GenomeZipper – a bioinformatics approach to unravel highly complex cereal genomes*
Betreuer: Prof. Dr. H.-W. Mewes
- 2001 – 2007 **Diplom Bioinformatikerin (Dipl.-Bioinf. Univ.)**
Ludwig-Maximilians-Universität & Technische Universität München
Titel: *Die Rolle der Stabilität der mRNA-Sekundärstruktur in der Genexpression*
Betreuer: Prof. Dr. W. Stephan
- 2000 – 2001 **Informatikstudium**
Ludwig-Maximilians-Universität
Studienfachwechsel zu Bioinformatik nach zwei Semestern
- 1994 – 1999 **Diploma de bacalaureat (rumänisches Abitur)**
Colegiul pedagogic A. Saguna, Sibiu, Rumänien

Berufspraxis

- 2007 – 2014 **wissenschaftliche Mitarbeiterin**
Helmholtz Zentrum München, Deutsches
Forschungszentrum für Gesundheit und Umwelt
Institut für Bioinformatik und Systembiologie
Ingolstädter Landstrasse 1, 85764 Neuherberg